From genomic variation to personalized medicine

Wesolowska, Agata; Schmiegelow, Kjeld

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

From genomic variation to personalized medicine

Agata Wesołowska-Andersen

14^{th} December, 2012

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions we know as disease.
Sir William Osler (1849-1919)

Preface

This thesis was prepared at the Center for Biological Sequence Analysis (CBS), Department of Systems Biology, at the Technical University of Denmark (DTU) in partial fulfilment of the requirements for acquiring the Ph.D. degree. The Ph.D. was funded by the Childhood Cancer Foundation, Danish Cancer Research Foundation (KB) and DTU.

All the work was carried out at the Center for Biological Sequence Analysis under the supervision of Associate Professor Ramneek Gupta, Professor Søren Brunak and Professor Kjeld Schmiegelow from Rigshospitalet.

Contents

Abstract

Genomic variation is the basis of interindividual differences in observable traits and disease susceptibility. Genetic studies are the driving force of personalized medicine, as many of the differences in treatment efficacy can be attributed to our genomic background. The rapid development of next-generation sequencing technologies accelerates the discovery of the complete landscape of human variation. The main limitation is not anymore the available genotyping technology or cost, but rather the lack of understanding of the functionality of individual variations. Single polymorphisms rarely explain a considerable amount of the phenotype variability, hence the major difficulty of interpretation lies in the complexity of molecular interactions.

This PhD thesis describes the state-of-art of the functional human variation research (Chapter 1) and introduces childhood acute lymphoblastic leukaemia (ALL) as a model disease for studying pharmacogenomic effects (Chapter 2 and 3). Chapter 4 describes the current interpretations of variations' effect and deleteriousness, accompanied by investigations of amino acid mutability compared to their deleteriousness presented in Paper I. Chapter 5 describes a pipeline used for calling variants from next-generation sequencing data and describes the common challenges encountered during analysis. Chapter 6 provides the motivation for a hypothesis-driven SNP selection and describes the publicly available resources used for this task. Following a review of the available large-scale genotyping techniques, Paper II introduces a novel cost-effective method for genotyping of a large custom SNP panel by means of multiplexed targeted sequencing and includes recommendations for efficient capture bait design. In Chapter 7 various methods of integrative analyses of genomic variations are presented, including testing of overrepresentation of rare variants, effects of multiple SNPs acting in the same biological pathway, contribution of coding variation to individual's disease risk, as well as identifying groups of patients differing in disease mechanisms defined by aberrations in protein-protein complexes. Chapters 8, 9 and 10 contain three papers applying the methods presented in Chapters 5 - 7 to investigate the heterogeneity of treatment response (Paper III), risk of infections (Paper IV) and disease aetiology (Paper V) in childhood ALL patients. Chapter 11 summarizes the thesis and includes some final remarks on the perspectives of genomic variation research and personalized medicine.

In summary, this thesis demonstrates the feasibility of integrative analyses of genomic variations and introduces large-scale hypothesis-driven SNP exploration studies as an emerging alternative to data-driven genome-wide association studies. Finally, the findings of the presented studies set new directions for future pharmacognenetic investigations and provide a framework for future implementation of personalized medicine.

Dansk resumé

Genomisk variation er årsagen til individuelle forskelle i fænotype, samt sygdomsmodtagelighed. Genetiske undersøgelser er den drivende kraft indenfor skræddersyet medicin, da mange forskelle i behandlingseffekt kan tilskrives vores egen genomiske baggrund. Den hurtige udvikling af næste generation sekventerings (NGS) teknologier accelererer opdagelsen af det fuldstændige landskab af den menneskelig variation. Den største begrænsning er ikke længere tilgængeligheden af genotypebestemmelses teknologi eller dens omkostning, men snarere den manglende forståelse af funktionaliteten af de individuelle variationer. Single nukleotid polymorfier (SNPs) kan sjældent forklare den betydelige mængde af fænotypens variabilitet, hvorfor det største problem i fortolkningen ligger i de komplekse molekylære interaktioner.

Denne ph.d.-afhandling beskriver den nuværende forskningsudvikling indenfor den funktionelle menneskelige variation (Kapitel 1), samt brugen af akut lymfoblastær leukæmi (ALL) hos børn som en modelsygdom til kortlægning af effekten af farmakogenomik (Kapitel 2 og 3). Kapitel 4 beskriver de nuværende fortolkninger af de molekylære effekter for genetiske variationer, ledsaget af undersøgelser af aminosyrernes foranderlighed, set i forhold til deres skadelige potentiale, hvilket præsenteres i Paper I. Kapitel 5 beskriver en arbejdsprocedure, som anvendes til at detektere varianter fra NGS data, samt de fælles udfordringer for analysen. Kapitel 6 beskriver både motivationen for en hypotese-drevet SNP udvælgelse, samt de offentlige tilgængelige ressourcer, der anvendes til denne procedure. Efter en gennemgang af de foreliggende store genotypebestemmelses teknikker, introducerer Paper II en ny omkostningseffektiv metode til genotypebestemmelse af en stor brugerdefineret SNP panel ved hjælp af multipleks målrettet sekventering, som indeholder anbefalinger for effektiv opsamling af bait design. I Kapitel 7 præsenteres forskellige metoder til integrationsfremmende analyser af genomiske variationer, herunder afprøvning af en overrepræsentation af sjældne varianter, virkningerne af forskellige SNPs som interagere i samme biologiske stofskiftevej, bidrag af kodningsvariation til en individuel sygdomsrisiko, samt identificering af patientgrupper med forskellige sygdomsmekanismer defineret ved aberrationer i protein-protein-komplekser. Kapitel 8, 9 og 10 indeholder tre artikler, der anvender de metoder, der præsenteres i Kapitel 5 – 7 til at undersøge heterogenitet af behandlingsrespons (Paper III), risiko for infektioner (Paper IV) og sygdomsætiologi (Paper V) i patienter med ALL. Kapitel 11 opsummerer afhandlingen og indeholder nogle afsluttende bemærkninger om udsigterne for genomisk variations forskning og personaliseret medicin.

Sammenfattende viser denne afhandling muligheden for integrationsfremmende analyser af genomiske variationer og indfører derved store hypotesedrevet SNP udforsknings undersøgelser som et spirende alternativ til datastyrede genom-dækkende associationsstudier. Endelig har resultaterne af de

fremlagte undersøgelser sat nye retningslinjer for fremtidige farmakogenomiske undersøgelser, samt skabt en ramme for fremtidige gennemførelser af skræddersyet medicin.

Acknowledgements

Work on this thesis would not have been possible without encouragement and support from many people. I would like to express my gratitude to my supervisors Ramneek Gupta, Søren Brunak and Kjeld Schmiegelow. Thank you all for sharing with me your motivation, enthusiasm, passion for research and immense knowledge throughout my PhD. I received from you a perfect combination of guidance and support, as well as plenty of freedom during these three years of research. Thank you for giving me the opportunity to be a part of many exciting projects and for creating an inspiring working environment.

I have been very fortunate to collaborate with many great people, including Louise Borst, Marlene Dalgaard, Bendik Lund, Susanne Rosthøj, Henrik Leffers and Martin Stanulla. None of the work presented in this thesis would be possible to achieve without your expertise, critical assessements and multitude of experimental work performed. It has been an extreme pleasure to work with all of you.

It has been a pleasure to be surrounded by many helpful people from CBS who always engaged in scientific discussions and provided me with many helpful insights. A special thanks to Laurent Gautier, Thomas Sicheritz-Ponten, Thomas Nordahl Petersen, Henrik Nielsen, Anders Gorm Pedersen and Simon Rasmussen.

Special thanks to the whole Functional Human Variation group, it has been a pleasure to share with you both the scientific interests at the weekly meetings, as well as the fun moments at our team-building events.

CBS has been not only a great place to work, but also to make friends. It has been a pleasure to share the office space with Josef, Tejal, Juliet, Kasper, Arcadio and Dhany. I had a lot of fun eating lunch together with you, sharing latest gossips over a cup of tea and pursuing our daily fruit hunts. And thanks to all other former and present colleagues for contributing to the friendly working environment.

The CBS system administration team has always been very helpful, I would like to thank John Damm Sørensen, Peter Wad Sackett, Kristoffer Rapacki and Olga Rigina for the technical support. The CBS administration never hesitated to help with any formalities, thank you for your help Lone Boesen, Dorthe Kjœrsgaard, Annette Vibeke Uldall and Martin Lund.

I would also like to thank all the people that agreed to comment on my thesis or its parts, especially Natasja Spring Ehlers, Rachita Yadav, Louise Borst and Kirstine Belling. Your comments were invaluable and helped to shape the final version of this thesis.

Finally, I would like to thank all my friends and family for their continuous support and encouragement. Especially Casper - thank you for always believing in me, being my best friend and my special person.

Papers included in the thesis

- Thomas Blicher, Ramneek Gupta, **Agata Wesołowska**, Lars Juhl Jensen, Søren Brunak. *Protein annotation in the era of personal genomics.* Current Opinion in Structural Biology, 20(3):335-341, 2010.

- **Agata Wesołowska***, Louise Borst*, Marlene Danner Dalgaard*, Laurent Gautier, Mads Bak, Nils Weinhold, Bettina Frydenlund Nielsen, Louise Rold Helt, Karine Audouze, Jacob Nersting, Niels Tommerup, Søren Brunak, Thomas Sicheritz-Ponten, Henrik Leffers, Kjeld Schmiegelow, Ramneek Gupta. *Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia.* Leukemia, 25(6):1001-1006, 2011.

- **Agata Wesołowska-Andersen***, Louise Borst*, Marlene Danner Dalgaard, Kirsten Kørup Rasmussen, Thomas Sicheritz-Ponten, Hans Ole Madsen, Hanne Vibeke Marquart, Claus R. Bartram, Peder Skov Wehner, Morten Rasmussen, Eske Willerslev, Torben Falck Ørntoft, Iver Nordentoft, Laurent Gautier, Søren Brunak, Martin Schrappe, Martin Stanulla, Ramneek Gupta, Kjeld Schmiegelow. *Extensive targeted SNP profiling predicts early treatment response and risk of relapse in 864 Danish and German childhood ALL patients.* Manuscript submitted to New England Journal of Medicine.

- Bendik Lund*, **Agata Wesołowska-Andersen***, Birgitte Lausen, Louise Borst, Kirsten Kørup Rasmussen, Klaus Muller, Helge Klungland, Ramneek Gupta, Kjeld Schmiegelow. *Variation in host genetics and infections during induction treatment in childhood acute lymphoblastic leukemia.* Manuscript ready for submission.

- Louise Borst*, **Agata Wesołowska***, Tejal Joshi, Rehannah Borup, Finn Cilius Nielsen, Mette Klarskov Andersen, Olafur G Jonsson, Peder Skov Wehner, Finn Wesenberg, Britt-Marie Frost, Ramneek Gupta, Kjeld Schmiegelow. *Genome-wide analysis of cytogenetic aberrations in ETV6/RUNX1-positive childhood acute lymphoblastic leukaemia.* British Journal of Haematology, 157(4):476–482, 2012.

* These authors contributed equally.

Papers not included in the thesis

- Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, Kivisild T, Zhai W, Eriksson A, Manica A, Orlando L, De La Vega FM, Tridico S, Metspalu E, Nielsen K, Ávila-Arcos MC, Moreno-Mayar JV, Muller C, Dortch J, Gilbert MTP, Lund O, **Wesołowska A**, Karmin M, Weinert LA, Wang B, Li J, Tai S, Xiao F, Hanihara T, van Driem G, Jha AR, Ricaut F-X, de Knijff P, Migliano AB, Gallego Romero I, Kristiansen K, Lambert DM, Brunak S, Forster P, Brinkmann B, Nehlich O, Bunce M, Richards M, Gupta R, Bustamante CD, Krogh A, Foley RA, Lahr MM, Balloux F, Sicheritz-Pontén T, Villems R, Nielsen R, Wang J, Willerslev E. *An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia.* Science, 334(6052):94 –98, 2011.

- Borst L, Buchard A, Rosthøj S, **Wesołowska A**, Wehner PS, Wesenberg F, Dalhoff K, Schmiegelow K. *Gene dose effects of GSTM1, GSTT1 and GSTP1 polymorphisms on outcome in childhood acute lymphoblastic leukemia.* J Pediatr Hematol Oncol, 34(1):38–42, 2012.

- Daniel Edsgärd, Marlene Danner Dalgaard, Nils Weinhold, **Agata Wesołowska**, Ewa Rajpert-De Meyts, Anne Marie Ottesen, Anders Juul, Niels E. Skakkebæk, Thomas Skøt Jensen, Ramneek Gupta, Henrik Leffers, Søren Brunak. *Genome-wide assessment of the association of rare and common copy number variations to testicular germ cell cancer.* Front Endocrinol (Lausanne). 2013;4:2.

- Marlene D. Dalgaard, **Agata Wesołowska-Andersen**, Nils Weinhold, Daniel Edsgärd, Søren Brunak, Anders Juul, Niels E. Skakkebæk, Ewa Rajpert-De Meyts, Henrik Leffers, Ramneek Gupta. *Identification of Genetic aberrations Associated to the Risk Alleles for Testicular Cancer.* Manuscript in preparation.

Part I

Introduction

Chapter 1

Genomic variation

With approximately 7 billion people living on earth, there are no two individuals that are the same. The differences between us are written in our DNA sequences. The completion of the Human Genome Project [127, 70] in 2001 began to shed lights on the details of our genetic code. Nowadays, with hundreds of individuals being sequenced every week, we begin to understand far more of the differences between us, however we are still far away from understanding the impact of every base in our 3-billion-base long genome. With every new generation, DNA is subject to mutation and recombination events, the sequences are mixed, shuffled, some bases are lost, added or turned around. These events result in many types of genomic variation, with the most important being:

- **Single nucleotide polymorphism (SNP)** is a single base difference in the DNA sequence occurring in at least 1% of the population[1]. It is estimated that SNPs constitute approximately 90% of all the human variation, and occur on average every 100-300 bases along the human genome comprising an estimated number of 15 million polymorphisms [5]. SNPs arise as point mutations and due to natural selection become fixed in the population.
- **Copy number variation (CNV)** is a form of structural variation where a section of DNA differs in number of copies between individuals. CNVs affect segments of sizes ranging from 1 kilobase to several megabases and can present as deletion, duplication, segmental duplication or inversion of the segment [4]. Copy number variations affect approximately 12% of the human genome [105].

[1]The exact definition varies depending on the source, however 1% seems to be the widely accepted definition

- **Variable number of tandem repeats (VNTR)** are patterns of repeating sequences of 2-60 bases. Depending on the length of the repeat the VNTRs are classified into micro- and minisatellites with up to 6-bases and above 6-bases repeated blocks respectively. VNTRs are heritable and are often used in forensic analyses or to analyse pattern of chimerism after haematopoietic stem cell transplantation.

- **Epigenetics** describes variation beyond the changes in the DNA sequence influencing gene expression. This mostly includes DNA methylation and histone modifications. DNA methylation is heritable and remains stable through cell divisions and may also last for multiple generations, while heritability of histone modifications is unknown [106].

Since the completion of the Human Genome Project in 2001 [127, 70] there have been several international efforts to catalogue human variation. Among them the International HapMap Project [45] and the 1000 Genomes Project [119] made significant contributions to our understanding of the common patterns of human genetic variation and the extent of variation between populations. Providing publicly available catalogues of common human variations greatly facilitated research of the genetic bases of susceptibility to different diseases and provided a foundation for the genome-wide association studies (GWAS). Data on genetic variation is available from the Single Nucleotide Polymorphism Database (dbSNP) [118] which serves as a central, public repository for genetic variation, where each variation has a unique reference identifier (rsID).

1.1 Genotype to phenotype

The consequence of the genetic variation is the variability in phenotypes, i.e. observable traits of organisms such as morphology, development or behaviour. The concepts of genotypes, phenotypes and their relationship have been described already in 1911 by Wilhelm Johannsen based on his observation of self-fertile common bean [61]. The simplest case of genotype to phenotype consequences are Mendelian traits which are controlled by a single locus and are characterized by a simple Mendelian inheritance pattern summarized by Mendel's laws [82]:

- **Law of Segregation.** Assuming diploidy, every individual has two alleles for a particular trait, and during formation of gametes the allele pairs segregate and each parent passes one randomly selected allele to the offspring. The trait in the offspring is dependent on the combination of dominant or recessive alleles.

- **Law of Independent Assortment**. Alleles of different genes assort independently of one another during gamete formation.

Those two laws were formulated by Gregor Mendel in 19th century based on his observations of colour, shape and position of the offspring of several generation of pea plants subjected to cross-hybridizing experiments. In humans genotypes determine traits like hair or eye colour, freckling, blood group, lactose intolerance or even earwax type. The latter is actually a Mendelian trait and is determined by a single SNP rs17822931 in *ABCC11* gene and can also be used to determine ancestry as the dry earwax is predominant among East Asians (80-95%) but seen very rarely (0-3%) in European or African populations [135]. Apart from physical traits and ancestry, genomic variation influences our susceptibility to various diseases. Mendelian disorders like Huntington's disease, cystic fibrosis or sickle-cell anaemia are caused by genetic aberrations in a single gene and are highly heritable. Because of high penetrance of the risk alleles, most of those genetic disorders are rare and are prevented by natural selection. Up to date more than 14,000 genes involved in more than 3,500 Mendelian disorders have been described and are collected in the Online Mendelian Inheritance in Man (OMIM) database [50].
However, majority of diseases and traits are influenced by a combination of genotypes from various associated genomic loci, which makes discovering the genetic determinants of a disease a rather challenging task. Furthermore, even though the genetic component has a major contribution to an organism's phenotype, the final result is usually influenced by the interplay between the genotype and the environment. This particularly well refers to complex disease susceptibility, where barely having a set of risk alleles at associating loci predisposing to a given disease is not a diagnosis by itself and an appropriate lifestyle can in many cases prevent the occurrence of the disease. A scientifically baffling example is height, believed to be a complex interplay of genetics, diet and environment.

1.2 GWAS

In recent years genome-wide association studies (GWAS) have become a popular way of exploring genomic variations related to specific traits or disease risk. The design of these studies includes collecting DNA samples from a large group of affected (cases) and unaffected individuals (controls). The samples are genotyped for usually around 500,000 - 1 million SNPs and then for each locus the differences in minor allele frequency (MAF) of the SNP are investigated between cases and controls. A p-value of the significance of this difference is calculated with a chi-squared test, and further corrected for multiple comparisons usually with Bonferroni correction [26].

Up to date more than 1,400 GWAS studies have been published, reporting more than 7,000 SNPs associated to various traits (Figure 1.1) [56]. The largest to date GWA study was conducted by the Wellcome Trust Case Control Consortium investigating 14,000 cases of common diseases and

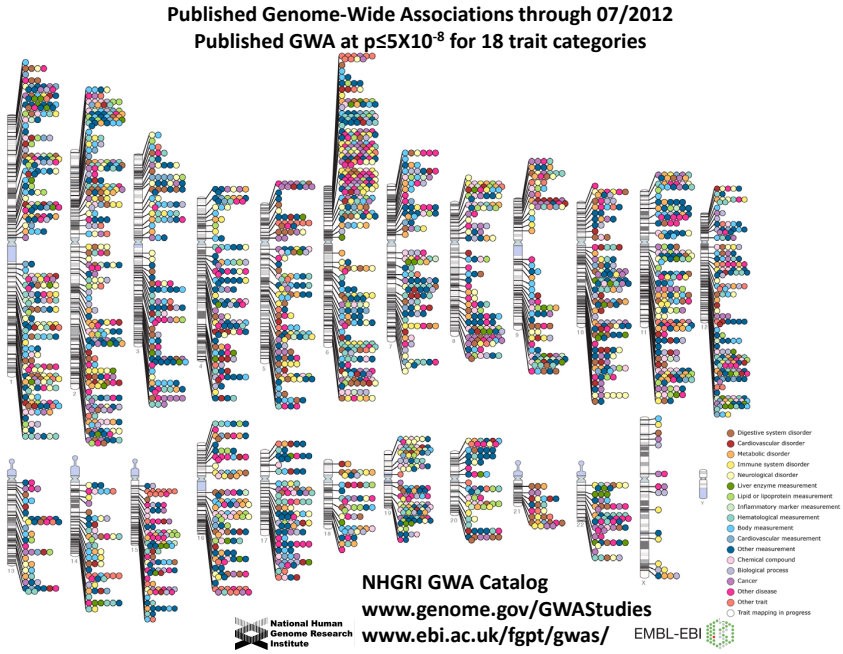


Figure 1.1. Findings of all Genome-Wide Association studies published up to July 2012 gathered across the chromosomes. Colour of the dots reflects the type of trait or disease investigated. Source: `www.genome.gov/gwastudies`.

3,000 shared controls [19]. This study demonstrated the power of GWAS to discover new disease genes and established the gold standard for the field. Despite the success of GWAS in detecting new disease associations, we are still far from translating the findings into clinical settings. The majority of SNPs investigated by commercially available genotyping arrays reside in non-coding regions of the genome, which poses a difficulty in the biological interpretation of the association signals. SNPs are inherited in linkage disequilibrium (LD) blocks, which means that there is little recombination activity within a block, resulting in similar MAF distribution among the SNPs within such block. In most cases the disease-associated SNP detected by GWAS is not the true causative SNP, but rather it is in high linkage disequilibrium with the causal variant. This creates a need for GWAS follow-up studies where the associated locus is investigated in detail to find the true causative variant [29]. An obvious deleterious candidate would be a SNP affecting the protein sequence and in turn protein function, however there exist also many regulatory variations affecting the expression of the gene or protein. These variants remain difficult to interpret as the existing annotations and our understanding of the function of the non-coding genome

is rather limited.
Another limitation of GWAS is that the assayed SNPs are selected to have considerably large MAF, which in turn leads to discovery of many common associated variations but all with rather small effect sizes. Despite many efforts and resources invested in investigating the common variations in human genomes, the promises to find the genetic components of heritable traits have lead to many disappointments. Even with regard to traits like height, with estimated heritability of 80-90%, the more than 200 associated genetic variants discovered up to date together account for approximately 10% of the trait heritability [3]. The missing heritability could be potentially explained by other factors including rare variants, copy number variations, epigenetics and environmental exposures influencing the final phenotype. This motivates investigating the heritability of traits and susceptibility to diseases applying systems biology approaches, where one should investigate the system as a whole, or at least at higher levels of functional abstraction, rather than concentrating on the contributions of the individual common variations.
Finally, the GWA studies investigate the genetic variations across the entire genome and present a data-driven study design. This creates a need for strict statistical corrections for multiple testing, as simply by chance many false positive results would be expected when testing thousands of loci. Despite the undoubted need for this procedure, many true weaker associated loci are discarded in this step. Another implication of the need to perform multiple testing correction is the need for large cohort sizes in order to obtain sufficient power to detect the associations with significantly low p-values. As a guideline, sample sizes of at least 1,000 cases and 1,000 controls are required to detect odds ratios ~1.5 in size with at least 80% power [138].

1.3 Identifying disease variants with NGS

An emerging alternative to GWAS is the next-generation sequencing (NGS) technology, which allows investigating of all the bases in the genome, including rare variants, indels and structural rearrangements. New sequencing technologies are developed constantly increasing the output and its quality, and at the same time decreasing the per-base cost. This leads to routine human genome sequencing becoming both feasible and affordable. With the massively parallel sequencing technologies we are able to sequence the whole human genome within a few days at a relatively low cost of approximately $10,000 per genome (source: `http://www.genome.gov/sequencingcosts/`), or even $5,000 from vendors such as Complete Genomics (`http://www.completegenomics.com/`). Among the most important applications of NGS are variant discovery, new genome assemblies, transcriptomics, methylation profiling and discovery of new microorganisms from environmental samples by metagenomics (Figure 1.2) [83].

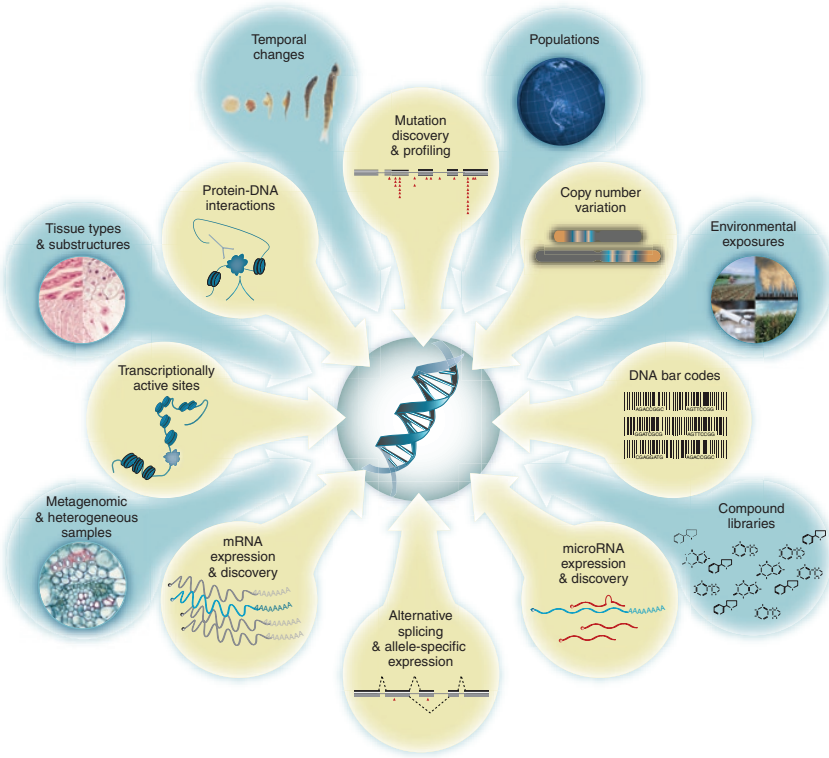


Figure 1.2. Applications of next-generation sequencing technologies. Source: Kahvejian *et al.* [64].

With the constantly decreasing cost of sequencing, NGS is likely to substitute SNP arrays for genotyping purposes in the near future [129]. The limitating factor in genome research is not anymore the available technology or its cost, but rather our abilities of interpretation of the genomic information. Even though we are able to sequence the whole human genome, our understanding is mostly limited to the protein coding regions constituting approximately 1% of the whole sequence corresponding to approximately 30 megabases (Mb) in length. For this reason exome sequencing became a popular strategy to identify disease-causing variation, as variations affecting protein sequence often result in loss of function of the protein and are therefore easy to interpret biologically. Exome sequencing is based on target enrichment technique where subset of whole genome DNA is captured by means of complementary RNA baits or a microarray and sequenced instead of the initial sample. This strategy proved to be extremely useful in identifying causal rare variants for Mendelian disorders, where in most cases the causative variants are non-synonymous coding with large effect

sizes [88, 89]. Exome sequencing has a significant advantage over microarray genotyping platforms as genotyping is not limited by probe design and it allows for detection of novel variants. Target enrichment can also be used on smaller custom genomic targets defined by needs of specific project and then multiplexing techniques can be used to further reduce the cost of the experiment. As compared to whole genome sequencing (WGS), targeted sequencing requires less sequencing output to produce required coverage and therefore more samples can be assayed at the same cost. An obvious limitation of exome sequencing is that it does not take into account any non-coding regions, which comprise 99% of the genome and might also contribute to disease risk. Additionally, WGS allows for examining the whole spectrum of genetic variants, including structural variations and copy number variations, while targeted sequencing is mostly useful for detecting SNPs and indels.

1.4 Personal genomes

Several companies took advantage of the advances in SNP genotyping technologies and published research findings of GWA studies and offer direct-to-consumer genetic testing of thousands of SNPs. Companies like 23andMe (`www.23andme.com`), deCODEme (`www.decodeme.com`) or Navigenics (`www.navigenics.com`) can analyse genetic variation for relatively low cost and report back the individual's risk for a number of common diseases, the ancestry, as well as some of the individual's traits. 23andMe has been aggressively marketing their product and the standard offering is a 1M Illumina Chip based SNP assay. The reported relative disease risk is evaluated by comparison with a disease risk of someone of the same age and gender in the general population. The health report can be a way of identifying diseases to which one is susceptible without the necessity of going to a doctor, however the results should be taken with a pinch of salt as the massive-scale SNP assays are not error free, and even though the overall percentage error might be very small, such errors could produce wrong risk assessment and produce a false sense of security or needless concerns. Even more concerning is the limited current understanding and predictability of most diseases. The findings from GWA studies of common diseases often have very low odds ratio and alone contribute very little to development of the disease. For instance, my own 23andMe report states that I have a 1.94x higher risk than average of developing Crohn's disease, which translates to 0.9% overall estimated risk. This report is not likely to raise my concern about my susceptibility to Crohn's disease. On the other hand, my risk of obesity according to my DNA is typical for people of my age, gender and ancestry, translating to my individual risk of 67.2%, which makes me more aware of the importance of diet and exercise despite the genotypes. Clearly, the opportunity to know what we are susceptible to can influence our life-style choices and in this way benefit our health, however there is also a possibility that a report of a high

risk of developing a certain disease can cause unnecessary anxiety in an individual. Often much more informative for the patient would be to investigate the family history and the environmental and life-style factors, which are likely to have much more influence on the disease risk than common genomic variation. On top of that, even though the genotyping concordance between the different platforms used by the three aforementioned companies are very high, the predicted genetic disease risks can be quite different [59], which leads us to believe that our understanding of the genetic basis of the diseases is actually very little and such information cannot yet be used in reliable assessment of the genetic contribution to disease risk of an individual.
Investigating individual genome variation can be used in a variety of ways beyond looking at disease susceptibility. There have been several successful published stories of defining the migration history and physical traits of ancient individuals, including the Saaqaq genome of an individual from the extinct Palaeo-Eskimo Saqqaq culture sequenced from a lock of hair preserved in permafrost [103] or the genome of an Aboriginal Australian sequenced from a lock of hair found in a museum [102]. Genome sequencing also tried to explain more contemporary questions, when the genome of a heavy metal rocker Ozzy Osbourne was sequenced in a hope to reveal the secrets of apparent lack of influence of his excessive drinking, drug abuse and partying on his health. Despite discovering a rare variant in a gene *ADH4* responsible for alcohol metabolism and variants pointing to higher likelihood of alcohol and cocaine addiction, the mystery remains unresolved as our understanding of the variations function is not sufficient to interpret them reliably. The Personal Genome Project [25] is currently trying to overcome some of those limitations, and improve our understanding of the ways how the genomic profile together with environmental exposures ultimately lead to traits. The study aims to recruit up to 100,000 individuals and to collect extensive information on their genomic sequence, tissues, environment, traits and others and undoubtedly will provide researchers with plenty of valuable data.

The development of modern medicine has almost exclusively been empiric without prior knowledge of the interactions between drugs and biological pathways. Not surprisingly, this frequently has lead to treatment failure or unacceptable toxicities.The work presented in this thesis aims at discovering the impact of genomic variations on treatment response and disease heterogeneity by combining the effects of variations in functional modules they are likely to operate in, defined by protein-protein interactions and biological pathways.

Chapter 2

Childhood acute lymphoblastic leukaemia

Acute lymphoblastic leukaemia (ALL) is a cancer of lymphoid cells, a subtype of white blood cells involved in the body's immune system. Acute refers to the relatively short time course of the disease, ALL would be fatal within few weeks if left untreated. The disease is characterized by excess of lymphoblasts, which are immature lymphoid cells not capable of performing their normal function, e.g. fighting infections. ALL arises from a malignant degeneration in a single lymphoid stem cell, which followed by dysregulated proliferation leads to clonal expansion and accumulation of lymphoblasts, which impairs normal haematopoiesis (Figure 2.1). Underrepresentation of normal erythrocytes, leukocytes and platelets in blood and bone marrow leads to clinical symptoms including anaemia, fatigue, pallor, bone pain, bruising and susceptibility to infections.

2.1 Epidemiology and aetiology

ALL is the most common malignancy affecting children, representing 25% of all paediatric cancers. The annual incidence of ALL is approximately 4 cases per 100,000 children in the Nordic countries, with peak incidence in children aged 2-5 years [57]. The causes of leukemia remain largely unknown, however genetic lesions leading to lymphoid stem cell transformation are believed to be triggered by a combination of environmental exposures, infections and inherited susceptibility [39].

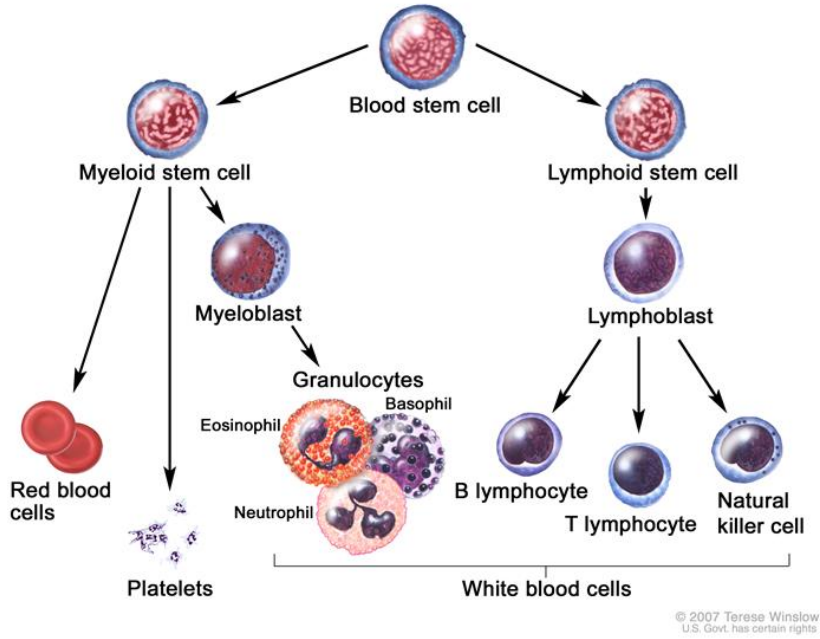


Figure 2.1. Blood cell development. Different blood and immune cell lineages, including T- and B-lymphocytes, differentiate from a common blood stem cell. A malignant degeneration in a single lymphoid stem cell may lead to development of ALL. Source: `www.cancer.gov`

2.2 ALL classification

There exist many subtypes of ALL characterized by different prognostic profiles. The subtypes reflect the lineage of lymphoid development affected, with approximately 85% of childhood ALL resembling B-cell lineage and 15% resembling T-cell lineage (Figure 2.1). Further classifications include specific cytogenetic aberrations, including abnormal number of chromosomes or chromosomal translocations. The most common aberrations in B-lineage ALL include hyperdiploidy, *t(12;21)* chromosomal translocation resulting in *ETV6/RUNX1* fusion gene and *t(9;22)* resulting in *BCR/ABL* fusion gene, while activating mutations in *NOTCH1* are the most common aberrations in T-lineage ALL [123] .

Childhood ALL patients from Nordic countries are treated according to common treatment protocols established by the Nordic Society for Paediatric Haematology and Oncology (NOPHO). At diagnosis patients are classified as standard risk (SR), intermediate risk (IR) or high risk (HR) patients. In the Danish cohorts studied in Papers III, IV and V this risk stratification is

based on 1) age at diagnosis, 2) white blood cell count (WBC) at diagnosis, 3) immunophenotype and 4) cytogenetics. Furthermore, minimal residual disease (MRD) levels at the end of induction therapy has been registered in the NOPHO ALL2000 protocol, but not applied for risk grouping, except for the very rare patient cases with >0.1% MRD levels after three months of therapy, who were offered haematopoietic stem cell transplantion in first remission [116].

2.3 Chemotherapy

The length of chemotherapy treatment in ALL is between 24 and 36 months depending on risk stratification and involves in the current NOPHO protocol up to 15 chemotherapeutic drugs, some of which are even given at varying doses or routes of administration. Treatment is divided into three phases [115]:

- **Induction phase** is a short and intensive treatment phase with the aim to kill most of the leukemic cells and obtain clinical remission, where no leukemic cells are detectable in bone-marrow aspirated by conventional morphological examinations. In this phase a combination of three to four drugs is used including a glucocorticosteroid, Vincristine (an antimicrotubule agent), an anthracycline and/or asparaginase.

- **Consolidation phase** is targeted at killing most of the remaining leukemic cells by alternating cycle of drug combinations, including antimetabolites, alkylating agents and an epipodophyllotoxin.

- **Maintenance phase** is a long and less intensive treatment phase. The aim of this therapy is to kill any residual leukemic cells, not eradicated by the induction and consolidation therapy to prevent relapse. Maintenance therapy is based on combination of following antimetabolites: mercaptopurine and methotrexate.

- In addition, **central nervous system directed therapy** is given.

The specific drugs used in childhood ALL treatment, together with their mechanisms of action, biology and relevant pharmacogenetics are described in Chapter 3.2.

Current cure rates for childhood ALL after first-line therapy approach 80-85% in the developed countries [116, 99]. However, even within risk-adapted treatment groups, there is substantial interindividual variability in treatment response. Nearly all chemotherapeutic drugs have narrow therapeutic range, which means that the differences between toxic and therapeutic

doses are small. Even though treatments are traditionally adjusted to patient's age, weight and other clinical parameters at presentation, still the interidividual differences in pharmacokinetics ("what the body does to the drug") and pharmacodynamics ("what the drug does to the body") largely influence the treatment efficacy. Many childhood ALL patients suffer from treatment-related toxic side effects, which are likely to be a result of too high bioavailability of a certain drug caused for instance by inefficient efflux system. On the other hand approximately 10-15% of patients experience a relapse of the disease, which in many cases could be attributed to insufficient drug disposition. Several studies have indicated that patients with the optimal host pharmacogenetics profiles have cure rates above 90% [33, 117]. Identifying failing mechanisms in the remaining patients and then individually adjusting the chemotherapy to mirror the drug disposition of the most favourable pharmacogenetics profiles could potentially bring the overall ALL cure rates to above 90%.

Chapter 3

Pharmacogenomics

Pharmacogenomics[1] is a particular field of studies dedicated to investigating genetic differences in metabolic pathways. Such variations are believed to affect the individual responses to drugs and determine the differences between effective and toxic drug doses. The pharmacogenetic polymorphisms can affect either the pharmacokinetics of the drug and reside then in genes involved in absorption, distribution, metabolism and excretion (ADME) properties of the drug, or the pharmacodynamics of the drug and involve variations in the drug targets and downstream signalling pathways [112].
The field of pharmacogenomics offers a great promise to the future of personalized medicine, where the drug dosing will be adjusted to patient's individual genetic makeup. Discovering the molecular mechanisms underlying drug exposure and effect will allow clinicians to minimize the toxic side effects and avoid the problem of lack of response due to too low dosage. A clear advantage of pharmacogenomics is that the genotype of an individual remains constant and is not affected by the treatment itself. Moreover, a variety of reliable genotyping methods exist which can be suited to the needs of a particular experiment [79]. The constant increase in genomic information data available facilitates the clinical interpretation of findings, and resources reviewing and curating this knowledge like the Pharmacogenomics Knowledgebase (PharmGKB) [54] or DrugBank [131] create a comprehensive foundation for the personalized medicine.

[1] Precise definitions of pharmacogenetics and pharmacogenomics differ depending on the source, however there is a general consensus on the two terms being interchangeable.

3.1 Pharmacogenetics in ALL

Childhood acute lymphoblastic leukaemia is for several reasons a model disease for studying pharmacogenetics effects [24, 33]:

- It has an early onset and therefore only limited environmental exposures playing a role in the etiology of the disease.
- It is the most common childhood cancer.
- The clinical subsets defined by the clonal karyotype are well-described.
- It is a 'liquid' tumour and is therefore relatively easy to study through blood samples or bone-marrow aspirates.
- It is in general highly chemosensitive.
- Patients experience significant interindividual differences in treatment response and toxicities.
- It is almost uniformly treated within the collaborative groups allowing for use of patient samples from other cohorts for validation of findings.

The bioavailability of drugs depends largely on efficiency of transport of those drugs inside the cell, how they are metabolized, and finally how rapidly they are secreted from the cell (Figure 3.1). Uptake of drugs inside the cell is mediated by the solute carrier (SLC) family of membrane trasnport proteins, which can modulate drug levels within the body by regulating their absorption, distribution, metabolism, and elimination (ADME) [114]. The most important enzymes metabolizing childhood ALL drugs include enzymes from cytochrome P40 superfamily activating or inactivating the drugs (phase I enzymes) and glutathione S-transferases enzymes conjugating drugs with endogenous substances which makes them suitable for excretion (phase II enzymes)[15]. Phase I enzymes comprise mixed function oxidases, which act mostly in the liver and by oxidation, reduction, hydrolysis or cyclization convert a prodrug to a pharmacologically active compound, or can transform a nontoxic molecule into a poisonous one. Phase II enzymes interact with the polar functional groups of phase I metabolites by conjugating reactions and are usually detoxicating. Efflux of the drugs is mediated mostly by the ATP-binding cassette (ABC) transporters, with the most studied member being P-glycoprotein, also known as multidrug resistance protein 1 (*MDR1*) or *ABCB1*. This ATP-dependent drug efflux pump has a broad substrate specificity, it influences the drug accumulation and often mediates the development of resistance to anticancer drugs [47]. The pathways of drug metabolism and transport are very polymorphic, which results in high interindividual bioavailability of the drug and subsequently differences in treatment response. Many of the enzymes share the same substrates and many cancer agents are metabolized by the same enzymes which creates a complex interplay of the drug dosage, pharmacogenetics and treatment response.

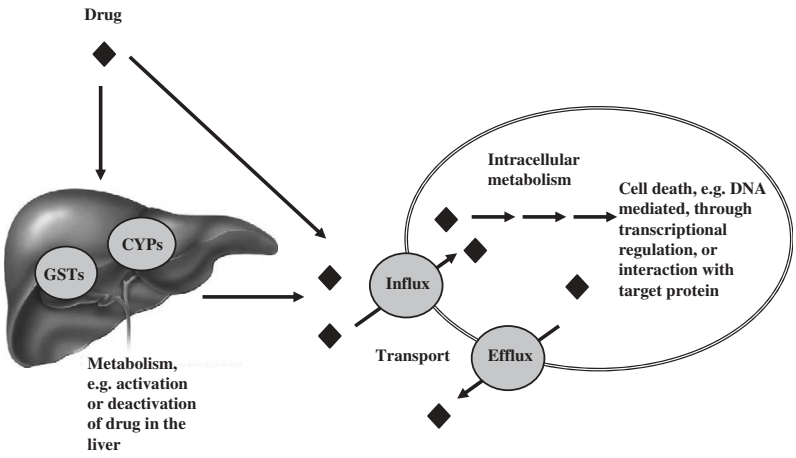


Figure 3.1. A simplified model of the general route of a drug with metabolism in the liver, transport in and out of the cell, intracellular metabolism, and final cytotoxic effect of drug. The arrows indicate direction of route. CYPs indicate cytochrome P450 enzymes; GSTs, glutathione S-transferases. Source of figure and legend: Davidsen *et al.* 2008 [33].

3.2 Drugs in childhood ALL

The following chapter briefly describes the chemotherapeutic drugs commonly administered during childhood ALL treatment, their mechanisms of action and the most important relevant pharmacogenetics domains.

3.2.1 Glucocorticoids

The glucocorticoid (GC) drugs commonly used in treatment of childhood ALL include prednisone, prednisolone (active metabolite of prednisone) and dexamethasone. They are metabolized in the liver by CYP3A and GST enzymes. Glucocorticoids act by binding to glucocorticoid receptor (*GR*) and by either binding to GREs (consensus sequence: GGT ACA NNNTGT TCT) of target genes or interacting with *AP-1* or *NF-κB* transcription factors they induce apoptosis in leukemic cells [125]. The response can be influenced by proteins involved in the *GR*-inactivating complex, including heat shock proteins 70 (*Hsp70*) and 90 (*Hsp90*), vitamin D receptor (*VDR*), and cytokines, such as tumour necrosis factor (*TNF*) or interleukins (ILs) [33].

3.2.2 Vincristine

Vincristine is an antimicrotubule agent exerting its anticancer effect by binding to tubulin, and thereby disrupting microtubule structures of the cell cytoskeleton and mitotic spindle leading to mitotic arrest and cell death [63]. Vinca alkaloids, including vincristine, are mostly metabolized by CYP3A enzymes, and are transported by several members of the ATP-binding casette transporters family. Among others *ABCB1*, *ABCC1*, *ABCC2*, *ABCC3*, *ABCC10* and *RALBP1* have been reported in association with vincristine resistance [94].

3.2.3 Anthracyclines

Two of the anthracycline drugs commonly administered in childhood ALL treatment are doxorubicin and daunorubicin. Anthracyclins interact with DNA by intercalation (squeezing between the base pairs) and inhibit replication processes by preventing progression of topoisomerase II (*TOP2A*) [85]. Anthracyclines are transported inside the cell by *SLC22A16* and exported by amongst others: *ABCB1*, *ABCC1*, *ABCC2*, *ABCG2* and *RALBP1*. The three main metabolic routes are: one-electron reduction (carried out by several oxidoreductases, including NADH dehydrogenases and nitric oxide synthases), two-electron reduction (carried out by various enzymes depending on the cell type, including *CBR1*, *CBR3*, *AKR1A* and *AKR1C3*) and deglycosidation (involving enzymes such as: *POR*, *XDH* and *NQO1*)[124].

3.2.4 Asparaginase

Asparaginase is an enzyme converting asparagine to aspartic acid and ammonia. In general, leukaemic cells do not synthesize asparaginase like normal cells, and are therefore dependent on its exogenous sources for survival. By catalysing the depletion of circulating asparagine, asparaginase leads to leukemic cell death [17]. Induction of asparagine synthetase (*ASNS*) in leukemic cells could potentially lead to asparaginase resistance [100].

3.2.5 Methotrexate

Methotrexate (MTX) is an anti-metabolite acting by inhibiting the dihydrofolate reductase (*DHFR*) and thereby inhibiting DNA synthesis and cellular replication by restricting access to folate coenzymes. It acts specifically during the S-phase of the cell cycle, where it prevents the growth and proliferation of dividing cancer cells. Pharmacogenetics of methotrexate is quite complex as the drug interferes with numerous components of the folate pathway, including: *TYMS*, *MTHFR* and *MTHFD1*, and affects both thymidylate synthesis and purine *de novo* synthesis. Methotrexate absorption is mediated mostly by *SLC19A1* and *SLC46A1* [48], and it is pumped out of the cell by several ABC transporters. Inside the cell it is polyglutamated by *FPGS*, and this process can be reversed by *GGH* [100, 84]. Compared to monoglutamated

MTX, the long-chained MTX polyglutamates are retained intracellularly and have increased affinity for the target enzymes. Polymorphisms in any of the mentioned genes might affect the systemic exposure of the drug.

3.2.6 Mercaptopurine

Mercaptopurine (6-MP) is an immunosuppressive drug, which upon conversion to active nucleotide metabolites by hypoxanthine phopshoribosyltransferase (*HPRT1*) is incorporated into DNA and RNA. 6-MP inhibits purine nucleotide synthesis and metabolism, and exerts its cytotoxic effect on leukemic cells by causing DNA strand breaks during aberrant post-replication mismatch repair [100]. Inactivating pathways catalysed by xanthine oxidase (*XDH*) or the polymorphic thiopurine methyltransferase (*TPMT*) are competing with the synthesis of active metabolites [137]. Some of the methylated 6-MP metabolites, most notably the methyl-thioinosine monophosphate, are strong inhibitors of purine *de novo* synthesis and may thus enhance the incorporation of thioguanine into DNA [52]. The dependency between *TPMT* activity and the effective 6-MP dose has been demonstrated in childhood ALL patients [108] (Figure 3.2), and it is up to date the only example of clinical translation of pharmacogenetics studies into ALL treatment protocols [107, 116].

3.2.7 Cytarabine

Cytarabine is an anti-metabolite drug, which after being metabolized to cytosine arabinoside triphosphate gets incorporated into human DNA instead of the highly similar deoxycytidine. During the S phase of the cell cycle the drug damages the DNA and blocks the progression of cells from the G1 phase to the S-phase [21]. The implicated mechanisms of resistance to cytarabine include: inefficient cellular uptake due to low levels or activity of *SLC29A1*, reduced levels of activating enzyme deoxycytidine kinase (*DCK*), or increased levels of inactivating enzymes 5'nucleotidase (*NT5C*) or cytidine deaminase (*CDA*) [69].

3.2.8 Cyclophosphamide

Cyclophosphamide is an alkylating agent metabolized in the liver to phosphoramide mustard, which after attaching to the alkyl group of the guanine base of DNA forms cross-links between and within DNA strands preventing DNA from being separated for synthesis or transcription. Additionaly, it can induce mispairing of the nucleotides and introduce mutations in the DNA, which together with inhibition of replication lead to disruption of DNA function and cell death [131]. Metabolic transformation of cyclophosphamide is mediated mostly by the CYP enzymes: *CYP2B6*, *CYP2C9* and *CYP3A4*, and the detoxification mostly by *ALDH1A1*.

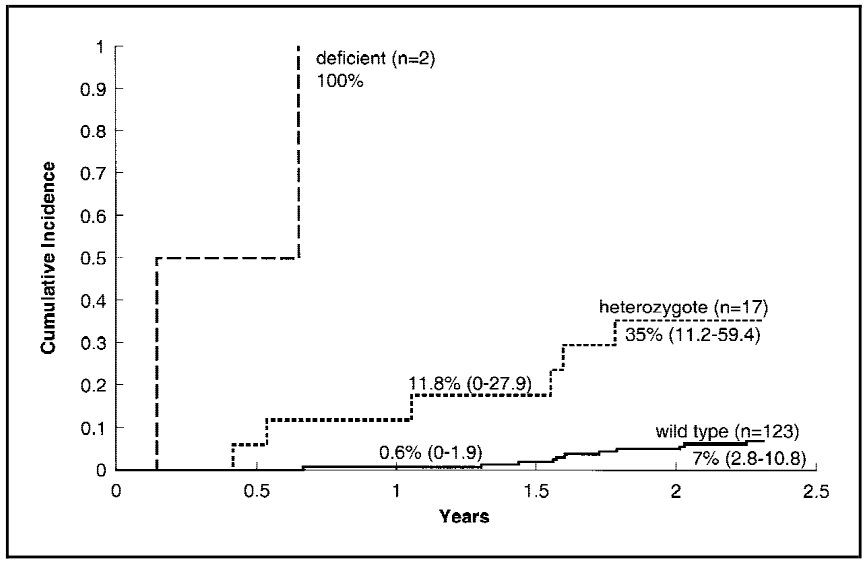


Figure 3.2. Cumulative incidence (95% confidence intervals) of requiring a decrease in 6-mercaptopurine dose (from 75 mg/m2 per day) to prevent toxicity among patients who were homozygous wild-type, heterozygous, and homozygous deficient for thiopurine S-methyltransferase (P<.001). Values = final cumulative incidences at the end of therapy and are also indicated at 1 year for those with heterozygous or wild-type status. Source of figure and legend: Relling *et al.* 1999 [108].

3.2.9 Epipodophyllotoxins

Two of the epipodophyllotoxins used in childhood ALL treatment are etoposide and teniposide, used mainly in the consolidation/intensification phases [40]. They act by inhibiting topoisomerase II enzyme (*TOP2A* and *TOP2B*), thereby preventing DNA re-ligation and causing breaks in DNA strands. The effects are cell cycle dependent and occur mainly during the S and G2 phases [51]. Etoposide is metabolized by *CYP3A* and *CYP3A5*, or it can be converted to O-demethylated metabolites by prostaglandin synthases (*PTGS1* and *PTGS2*) or myeloperoxidase (*MPO*). The metabolites can be inactivated by *GSTT1*, *GSTP1* and *UGT1A1*, and efflux is mediated by *ABCC1*, *ABCC3* and *ABCB1* [132].

Part II

Methods

Chapter 4

Predicting SNP effects

The use of exome and genome sequencing in disease genetics allows us to detect previously unknown variants in genomic samples. The biggest challenge of those studies is the interpretation of the variants and prioritizing them to dissect the causative variants from neutral variants from a list of thousands of polymorphisms. A successful approach has been developed for identifying the causative variants for Mendelian disorders using exome sequencing, where common variants are discarded and rare non-synonymous variants are considered to be the most likely candidates [88, 89]. This strategy however would not be successful in studying common diseases, therefore being able to predict the functional effect of any variation is crucial to interpret the impact on the susceptibility to disease.

4.1 SNP effect on transcript

As SNPs occur on average every 100-300 bases along our genome, we can expect that the vast majority of them would not have a functional effect on any protein. The effect of the SNP depends on its location relative to the coding sequence or regulatory elements. Examining the effect of a variation allele on a transcript allows for predicting the likely functional effect of the variation [81]. Different types of variation effects exerted on the transcripts are illustrated in Figure 4.1.

4.2 Protein-coding changes

The polymorphisms affecting protein-coding sequence are the most studied up to date, with the interpretation being guided by applying both evo-

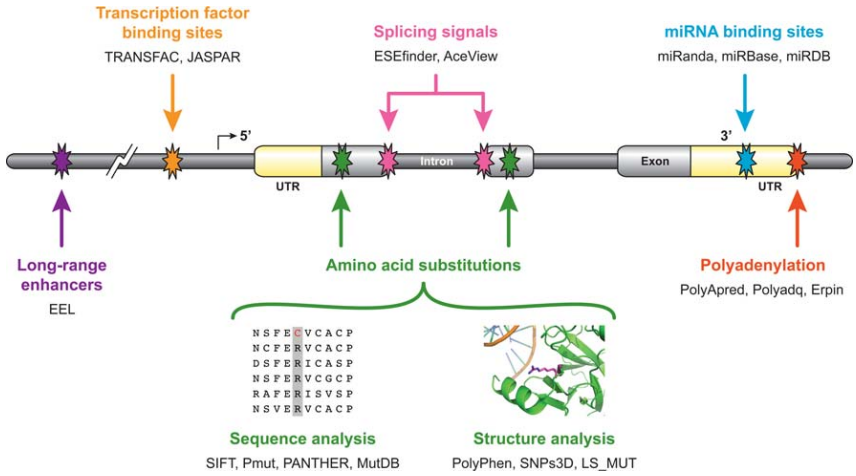


Figure 4.1. Consequences of variations relative to transcript structure together with computational tools which can be used for mapping or analysis of the sequence variants in each category. Source: Lee *et al.* 2009 [72].

lutionary and biochemical evidence. This class of variations is also best understood as any variation resulting in an amino acid change, premature stop codon or changing the coding frame is likely to affect the function of the protein. The strongest and most widely applied predictor of deleteriousness is the evolutionary conservation based on the assumption that the natural selection drives the evolution of species by eliminating deleterious mutations and conserving the essential positions. The impact of a mutation might also depend on the location of the variation in specific structural domain of the protein or a binding site, as well as on the similarity of the biochemical and physical properties of the substituted amino acids. Based on comparison of disease-causing non-synonymous coding SNPs from HGMD Pro 2009 [28] and all the non-synonymous coding SNPs from Ensembl release 54 [58] we demonstrated that the mutability of amino acids does not correlate with the deleteriousness of the change (Figure 1a in Paper I). Transitions between more similar amino acids are more permissible and substituting for amino acids differing in physicochemical properties is more likely to be disruptive for the protein (Figure 1b,c,d in Paper I). There exist many computational tools using these information for predicting whether a given non-synonymous coding variation is deleterious, including the most popular: SIFT [87] and PolyPhen-2 [1]. Excellent reviews of those and variety of other computational SNP effect prediction tools are provided by Lee *et al.* 2009 [72] and Cooper *et al.* 2011 [29]. While assessing the deleteriousness of a mutation can be very valuable, the ultimate goal is to understand what

changes a given variation exerts on the molecular level. For this purpose the EPipe Consortium (`http://www.cbs.dtu.dk/services/EPipe/`) has developed a pipeline for assessing differential predictions of functional changes between the wild-type and mutated proteins (Paper I). EPipe integrates a large number of individual prediction tools, including predictions of sub-cellular localization, post-translational modifications and sequence motifs, allowing for more fine-grained assessment of the impact of variation on the function of the protein.
Finally, it is important to consider that synonymous variations, even though they do not change the final protein sequence and are usually considered to be non-functional, have been demonstrated to also affect protein structure and function by changing the translation kinetics and affecting protein folding [66, 65]. Those instances may be not as frequent and obvious as the causality of non-synonymous coding variants, but they should not be ignored.

4.3 Non-coding variations

Since the protein-coding portion of the genome constitutes only 1% of its total length, the majority of genetic variations reside in the non-coding regions. Evolutionary analyses show that many non-coding regions are also conserved between species implying functional importance of those regions. Moreover, many GWAS results reside in non-coding regions and are not in linkage disequilibrium with any coding variations, which supports the hypothesis of functionality beyond coding regions. Apart from the coding region of the genome, there exist many regulatory variations which may reside in microRNA binding site, promoter region, transcription factor binding site or other regulatory feature, resulting in changed gene expression or alternative splice variant. It is estimated that any individual has many more functional regulatory variants than coding variants, however they are likely to have smaller effect sizes [128]. Similarly as for coding variants, the basis for assessing deleteriousness of non-coding variant is comparative genomics, however due to the rapid evolution of the non-coding DNA it includes only comparison of closely related species, and can only be applied to homologous regions, missing the human-specific sequence.
Several experimental efforts are addressing those issues, including the Encyclopedia of DNA Elements (ENCODE) Consortium attempting to generate whole-genome functional annotations, including non-coding regions like non-coding RNAs and *cis*-regulatory elements in different cell types [110] and a range of expression quantitative trait loci (eQTL) studies are detecting variations affecting gene expression. Regulatory variations might be more complicated to validate experimentally due to their spatial and temporal patterns of action.

Even though significant progress is being made in the field of functional human variation, we are still far from understanding the impact of any variation on the molecular level. An important consideration is that even proving

molecular functionality of a variation (e.g. loss of protein function or change of expression) does not always prove the deleteriousness of the variant or its causality with respect to the studied phenotype. The task of interpretation of genomic changes is complicated by the fact that at many positions in the genome there exist several overlapping transcripts, and a single variation can have different effects on any of them. To fully understand the functional result of the variation it is necessary to integrate into analysis the knowledge about tissue-specific gene expression and alternative transcripts, as well as details of the mechanisms of regulation. Finally, the function of each variation should be assessed in a broader context, investigating how it can affect protein-protein interaction networks or biological pathways.

4.4 Paper I- Protein annotation in the era of personal genomics

The following paper describes the current state of art of annotation of protein function and introduces an integrative pipeline for investigating differential predictions of protein function, as well as presents some perspectives for the personal genomics.
As a part of genomic variation background presented in the paper, we have collected data on the known coding variations at the time from Ensembl release 54 (NCBI genome build 36) and investigated their distribution across the genome (Figure 2, Paper I), as well as compared the mutability and deleteriousness of amino acid changes based on comparison of all known non-synonymous coding SNPs and the disease-causing SNPs extracted from HGMD Pro 2009 (Figure 1, Paper I).

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Protein annotation in the era of personal genomics

Thomas Blicher[1,2], Ramneek Gupta[2], Agata Wesolowska[2], Lars Juhl Jensen[1] and Søren Brunak[1,2]

Protein annotation provides a condensed and systematic view on the function of individual proteins. It has traditionally dealt with sorting proteins into functional categories, which for example has proven to be successful for the comparison of different species. However, if we are to understand the differences between many individuals of the same species — humans in particular — the focus needs be on the functional impact of individual residue variation. To fulfil the promises of personal genomics, we need to start asking not only what is in a genome but also how millions of small differences between individual genomes affect protein function and in turn human health.

Addresses
[1] NNF Center for Protein Research, Faculty of Health Sciences, Blegdamsvej 3b, DK-2200 Copenhagen N, Denmark
[2] Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Bldg. 208, DK-2800 Lyngby, Denmark

Corresponding author: Blicher, Thomas (Thomas.Blicher@cpr.ku.dk), Gupta, Ramneek (ramneek@cbs.dtu.dk), Wesolowska, Agata (agata@cbs.dtu.dk), Jensen, Lars Juhl (LarsJuhl.Jensen@cpr.ku.dk) and Brunak, Søren (blicher@cbs.dtu.dk)

Current Opinion in Structural Biology 2010, **20**:335–341

This review comes from a themed issue on
Sequences and topology
Edited by Sarah Teichmann and Nick Grishin

Available online 18th April 2010

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2010.03.008

Introduction — personal genomics is here

The era of personal genomics has begun. DNA sequencing technologies are constantly becoming faster and cheaper, and human genomes are already being sequenced by the thousands (1000 genomes consortium, http://1000genomes.org/; Personal Genome Project, http://www.personalgenomes.org/). However, without proper downstream analysis, genome sequencing is most useful for the understanding of chromosomal organization and stratification of individuals, and does not lead to the identification of functional variation. To harvest the benefits of personal genomes, it is not enough to sequence them — we also have to annotate and compare large-scale functional variation (Figures 1 and 2).

Protein annotation is an old discipline. Ever since the first protein sequences became available there has been a need for identifying their functional features, since not all amino acid residues in a sequence are equally important to the function of the protein. At a coarse-grained level, annotation is a matter of putting proteins into the right boxes, which can be formalized by ontologies describing protein structures, molecular functions, biological processes and signalling pathways. One example is Gene Ontology [1]. At the most fine-grained level, annotation aims at pinpointing specific functional residues that constitute, for example, the active sites of enzymes, post-translational modification sites or essential structural motifs.

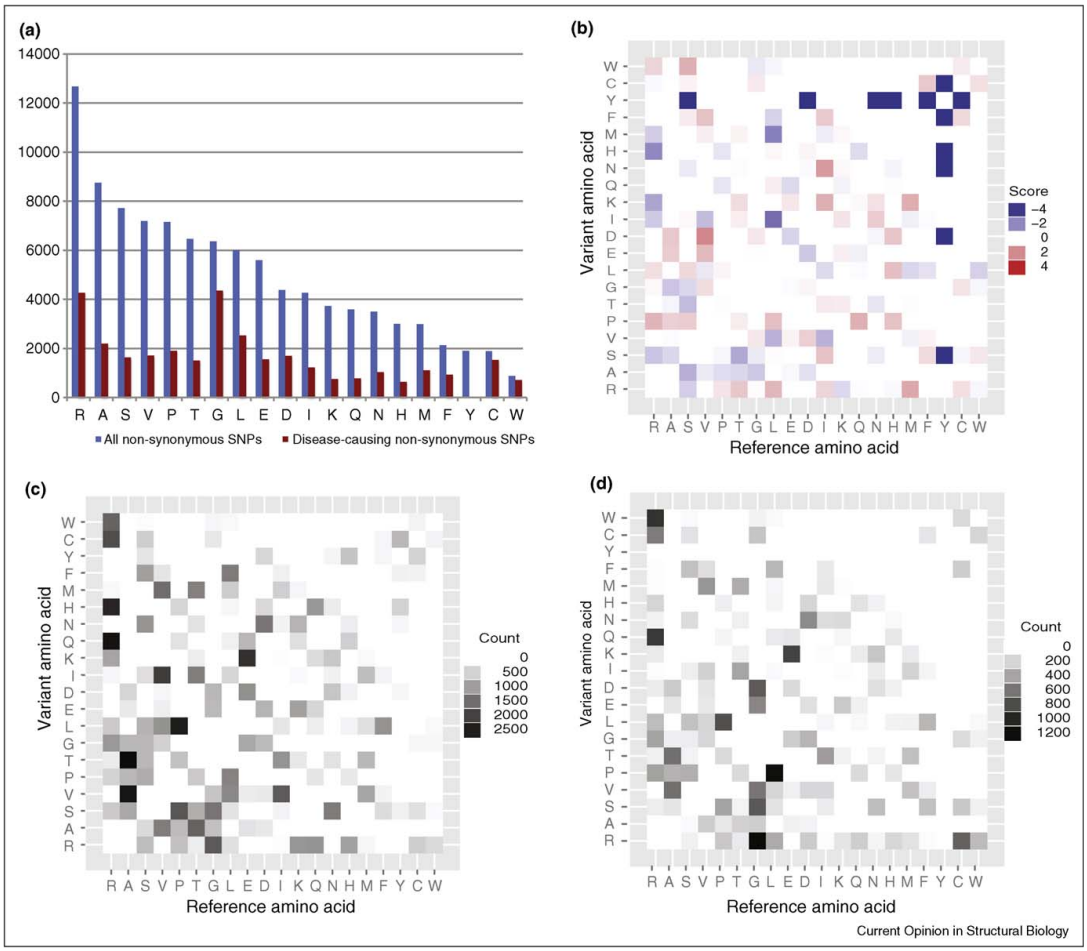
The overall goal of protein annotation is to provide a structured view on function to allow for a more efficient use of the available data. An important point of such efforts is that annotated data lends itself more easily to analysis, comparison and integration with other kinds of data. Automated annotation of structural and functional properties of proteins from their amino acid sequences is often possible, because similar functional or structural elements can be identified through homology (common ancestry) with already annotated proteins. In light of the flood of data being produced, automation of annotation is becoming increasingly important for complementing manual annotation [2]. This also means that the individual tools used for annotation — as well as more complex workflows integrating many such tools — must be able to deal with the rapidly growing amounts of data, and some newer prediction methods used for sequence annotation have indeed been cleverly designed to automatically update and retrain themselves as new data becomes available [3••].

Here, we review the current status of methods for the annotation of protein function as seen in light of the recent developments within the field of personal genomics. With the current data challenge in mind, we describe the different levels of annotation and which types of features to annotate, and summarize the most relevant annotation tools, namely predictors of local (positional) and global protein features. We then introduce some examples of more advanced workflows in which the outputs from multiple feature predictors are combined. Finally, we discuss some of the challenges involved in moving from single sequences and simple tools to pan-genome comparisons through complex workflows. As a special case we will focus on the promises of comparative protein annotation in the context of human health.

Sequence versus structure

Protein structure is commonly considered the main determinant of protein function, although the importance of

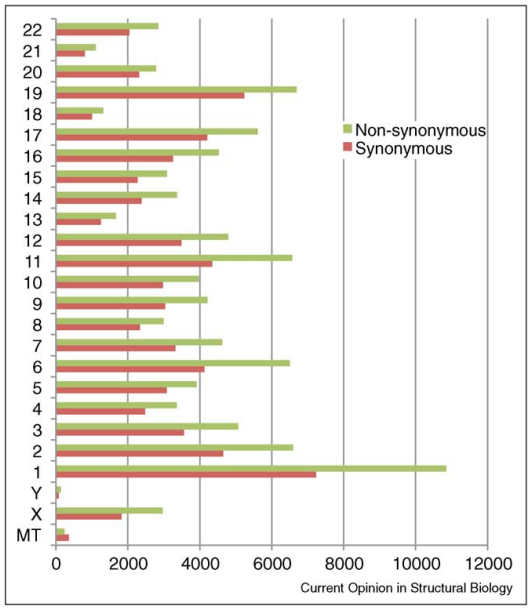
Figure 1



Human variation and disease-causing SNPs. Disease-causing coding SNPs often perturb structure and/or function in proteins. This figure looks at amino acid changes induced by disease-causing non-synonymous coding SNPs (35 445 SNPs from HGMD Pro 2009) compared to a background of all non-synonymous coding SNPs (100 410 SNPs from Ensembl release 54; NCBI genome build 36). Mutability of amino acids (barplot in **(a)**) shows Arg (R) to be the most changed while Cys (C) and Trp (W) to be the least changed. In this statistics, 70% of Gly (G) changes and 80% of Cys and Trp changes are disease-associated. Heatmaps show a comparison as log 2 ratio of normalised counts (**b**) of all non-synonymous changes (**c**) and disease-associated changes (**d**). Blue data points (b) represent more permissible amino acid transitions, while the red data points represent more disease-causing ones. Various symmetries can be seen with the blue data points reflecting the fact that amino acids similar in physicochemical properties can be interchanged without being disruptive (e.g. I and L, D and E, T and S) opposed to substituting for amino acids differing in properties (e.g. W and S). In case of disease-causing transitions (red data points), the symmetries are fewer, suggesting such functional changes are often disruptive and occurring unidirectionally. The transitions to and from tyrosine (Y) were not identified in the disease-causing SNPs dataset, possibly due to limited size of the data set, thus marking strong blue spots.

intrinsically unstructured regions is becoming increasingly recognized [4]. While sequence conservation usually implies structural similarity, the same is not always true for sequence–function relationships. This is illustrated by paralogues, which often have different functions despite having similar sequence and structure. Moreover, a single mutation can fundamentally change the function of a protein; for example, a mutation in a binding pocket can entirely alter the substrate specificity of an enzyme. Whereas inherited mutations with such

Figure 2



Chromosomal distribution of SNPs in coding regions. The plot shows the number of SNPs per chromosome (MT = mitochondrial). Over 71K synonymous and 100K non-synonymous SNPs in coding regions of the human genome are recorded in Ensembl release 54 (NCBI genome build 36). Chromosome sizes decrease from Chr-1 (247M) to Chr-22 (49M), but the number of non-synonymous SNPs generally correlates poorly with this. Especially Chr-17 (79M) and 19 (64M) have disproportionately high numbers of non-synonymous SNPs.

dramatic consequences are strongly selected against on an evolutionary time scale, they are highly relevant in the context of understanding variation between individuals. This means that overall sequence and structure similarity are generally useful for inferring function among well-conserved proteins, but carry little or no information about the functional consequences of random point mutations, whether they are inherited or somatic.

To overcome this challenge, the functionally important residues in the proteins must be identified, implying a change in focus from global to positional features. One way to do this is to manually annotate proteins or protein families with information on, for example, catalytic and ligand-binding residues as is done by SMART [5], the Catalytic Site Atlas [6] and Firestar [7]. Unfortunately, our knowledge on functionally important residues is very incomplete, which prevents comprehensive manual annotation. We must thus rely on predictions. Fortunately, many functionally important residues are part of so-called linear motifs, which are short, evolutionarily variable sequence patterns involved in post-translational modifications, protein targeting, cleavage and interactions. Due to their short length and rather poor conservation, most linear motifs are difficult to predict. Nonetheless, depending on the availability of experimental data prediction methods have been developed for many of the most important ones.

Prediction of linear motifs and protein features

Tools for predicting protein features come in different flavours, the main difference being whether they predict positional or global features.

The proper subcellular localisation of a protein inside eukaryotic cells is crucial for the function, as it determines the availability of other potentially interacting proteins and metabolites. Thus, many prediction tools deal with subcellular localisation, and some of the most successful methods are SignalP [8], Phobius [9], Philius [10], TargetP [11], WoLF PSORT [12], BaCelLo [13] and LOC-Tree [14]. Although subcellular localisation per se is a global property of the protein, the localisation signals that control it are most often positional. To understand differences in subcellular localisation between individuals one must therefore look at these positional motifs — global properties like amino acid composition will not be particularly informative.

Other types of protein features are much more readily identified as being controlled by the presence of motifs and thus allow for simpler interpretation in the context of protein variants. Such positional features include all the well-known post-translational modifications: Attachment of lipid moieties to direct proteins to the proper cellular location through anchoring to the cell membrane [15,16]; various kinds of glycosylation and glycation to help fold, protect and stabilise proteins [17–19]; phosphorylation, the most common post-translational modification and a crucial part of all intracellular signalling events [3••,20]; ubiquitination [21] and sumoylation [22] to regulate a wide range of protein functions. In this context, motifs for binding, recognition and peptide cleavage can also be considered easily interpretable.

Prediction of transmembrane protein segments and thus implicitly subcellular localisation has also been highly successful, where one of the best tools for predicting transmembrane helices in protein sequences is Philius [10]. Although it does not rely on the presence of specific motifs as such, it is guided by special amino acid preferences inside the transmembrane helices as well as at the helix boundaries. Prediction of other properties, for example globular protein secondary structure [23,24], surface exposure of individual amino acid side chains [25], destabilisation upon point mutations [26] and intrinsic protein disorder [27,28] also rely on such more distributed sequence features.

A number of tools are available for the prediction of deleterious single amino acid substitutions and typically combine homology considerations (alignment) with information derived from protein structure and physicochemical properties of the exchanged amino acids [29–32].

Complex workflows and pipelines

When analysing the effects of single point mutations, multiple sequence-based predictions need to be run on the same sequence and the results viewed in the context of each other. A few prediction pipelines, which automate this task, are currently available.

The Eukaryotic Linear Motif resource for functional sites in proteins, ELM [33•,34], contains a large number of regular expressions for identifying linear motifs. A number of context-based rules and filters are employed to improve predictive power and cover subcellular localisation, phylogeny (homology), protein structure and Pfam domain definitions [35]. The output from ELM is presented in an easily accessible form to allow for a quick overview of the individual predictions.

Dasty2 is a web-based service capable of running numerous prediction servers and annotation databases [36] complying with the distributed annotation system (DAS) standard [37]. The results from the remote servers are collected and presented to the user in an interactive and user-friendly format. Although Dasty2 and ELM produce superficially similar outputs, they actually represent two fundamentally different ways of creating annotation. While ELM stores all prediction tools locally and thus is an example of a centralised resource, Dasty2 relies on remote services and thus is distributed computing-wise.

Using tools such as ELM or Dasty2, it is relatively easy to arrive at consensus predictions and to gauge the different types of prediction against each other. However, neither method allows for any simple way of identifying predicted features, which are functionally differential, the central element of personal genomics. When studying genetic variation between individuals, be it single-nucleotide polymorphisms (SNPs) either coding or non-coding, larger structural variation within genes (insertions and deletions), somatic mutations or simply alternatively spliced variants of the same gene, it is absolutely essential to be able to zoom in on the differences rather than looking at all functional aspects. Such differences are likely to be responsible for the main changes in the behaviour of the protein.

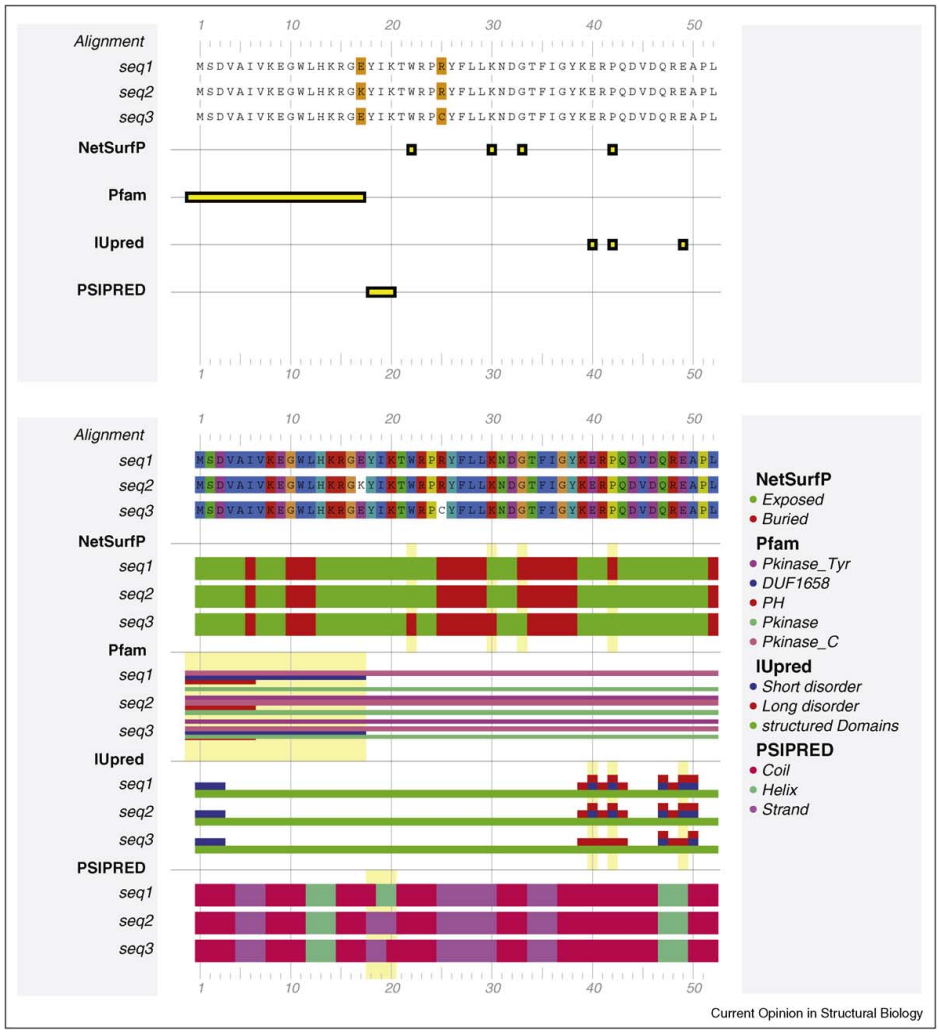
A pipeline focusing on functionally differential features is currently in development by the EPipe Consortium (http://www.cbs.dtu.dk/services/EPipe/). It is similar to both ELM and Dasty2 in the sense that EPipe integrates a large number of individual prediction tools, both local and distributed (including selected linear motifs from ELM). However, unlike both ELM and Dasty2, sets of related sequences can be submitted to EPipe for comparison, and differential predictions are highlighted in the multiple alignment of the sequences, and if available also on protein structures. An example of the kind of analysis facilitated by EPipe is shown in Figure 3.

To understand the impact of protein variation and ultimately the cause of disease-associated changes to proteins, tools for comparative sequence analysis are absolutely essential, as they facilitate the separation of inconsequential changes to the protein scaffolds (neutral changes) from changes affecting crucial protein function.

Discussion and outlook

In the above, we have argued that personal genomics requires a change in how we approach protein annotation. In particular, the focus must change from the functions of entire proteins to smaller functional building blocks to facilitate meaningful comparisons. However, one must not lose sight of the fact that proteins are social molecules, which interact with each other in many different settings. Thus, to fully appreciate the impact of small molecular changes such as mutations one needs to see them in the proper context. If, for example, a protein under the

Figure 3



Using EPipe to study AKT1 variants. E17K and R25C are opposing somatic mutations in AKT-1 affecting the pleckstrin homology domain; E17K leads to constitutive activation by pathological localisation to the plasma membrane and decreases sensitivity to an allosteric kinase inhibitor. R25C results in a kinase that does not efficiently bind phosphoinositides, fails to localize to the membrane, and is not activated. EPipe indicates that E17K induces changes in secondary structure (PsiPred), surface exposure (NetSurfP), and a Pfam domain 'DUF1658' found many times in the genome of *Coxiella* (known to exhibit antiapoptotic activity via AKT1) interestingly disappears [52]. R25C shows a change in surface exposure, secondary structure and disordered regions (IUpred). All of these predictions imply structural changes that, in the absence of other information, could prioritise novel mutations for further investigation. Only part of the alignments are shown in this figure.

control of a kinase loses a phosphorylation site, the overall impact should be analysed by looking at the interaction partners both upstream and downstream in the signalling pathway and not only by looking at the changed protein itself. Somatic mutations in cancer are often activating or deactivating a pathway, where variations across different members of a pathway eventually lead to the same phenotypic effect. A single point mutation can drive the activation of a pathway such as the E17K mutation in the AKT-1 pleckstrin homology domain (Figure 3) that forms one of several genetic mechanisms that can activate the PI3K/AKT cell-proliferation pathway [38].

In the case of kinase/phosphoprotein-driven signalling, such network considerations have already been formalized in the NetworKIN method [39•]. The impact of small molecular changes could also be studied in connection with protein complexes known to be involved in human diseases as demonstrated recently [40,41••]. Other pathway databases such as the Kyoto Encyclopedia of Genes and Genomes, KEGG [42], take an even broader view on things. Here, metabolic and regulatory pathways, drug interaction networks and disease associations can be studied from the molecular level all the way up to tissues and organisms. For an overview of pathway databases, see the Pathguide online resource described by Bader *et al.* [43].

Integrating pathways with an assessment of the effect of individual molecular changes would represent an important step forward in our general understanding of the impact of variation. As an example, take the case of allelic variants of cytochrome p450 enzymes, a large family of enzymes involved in general drug metabolism [44]. Understanding the functional impact of variations in such important enzymes using some of the approaches mentioned above could be useful in the context of achieving truly personalized medicine. Many SNP associations can be traced to coding changes in proteins such as blood group [45], skin colour [46], sickle-cell anaemia [47] and metabolic cold adaptation [48,49] to name to few. Phenotypic annotation of protein (and non-coding) variants still needs further work, but is already enabling efforts reconstructing features from anonymous DNA [50••,51••]. Several commercial efforts such as deCODEme, 23andMe, Knome, BioResolve and Navigenics are already taking a limited amount of individual variation-associated phenotypic predictions directly to consumers at an affordable price.

If personal genomics is to have an impact on our understanding of human biology and health, it is essential to put the small molecular variations among humans into the context of the pathways and networks they operate in and eventually relate to the individual instead of race or geographic region.

Acknowledgements

The authors wish to acknowledge the FP6 Biosapiens and EMBRACE grants, as well as support from the Villum Kann Rasmussen Foundation. The work carried out in this study was in part supported by the Novo Nordisk Foundation Center for Protein Research.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**:25-29.

2. Hinz U: **From protein sequences to 3D-structures and beyond: the example of the UniProt knowledgebase**. *Cell Mol Life Sci* 2010, **67**:1049-1064.

3. •• Miller ML, Jensen LJ, Diella F, Jorgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T *et al.*: **Linear motif atlas for phosphorylation-dependent signaling**. *Sci Signal* 2008, **1**:ra2.
Highly systematic and comprehensive description of protein phosphorylation networks. The resource uses phylogenetic trees to structure the currently available data and stays updated autmatically.

4. Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ: **Understanding eukaryotic linear motifs and their role in cell signaling and regulation**. *Front Biosci* 2008, **13**:6580-6603.

5. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5, domains in the context of genomes and networks**. *Nucleic Acids Res* 2006, **34**:D257-260.

6. Torrance JW, Bartlett GJ, Porter CT, Thornton JM: **Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families**. *J Mol Biol* 2005, **347**:565-581.

7. Lopez G, Valencia A, Tress ML: **Firestar — prediction of functionally important residues using structural templates and alignment reliability**. *Nucleic Acids Res* 2007, **35**:W573-577.

8. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**:783-795.

9. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method**. *J Mol Biol* 2004, **338**:1027-1036.

10. Reynolds SM, Kall L, Riffle ME, Bilmes JA, Noble WS: **Transmembrane topology and signal peptide prediction using dynamic bayesian networks**. *PLoS Comput Biol* 2008, **4**:e1000213.

11. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence**. *J Mol Biol* 2000, **300**:1005-1016.

12. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucleic Acids Res* 2007, **35**:W585-587.

13. Pierleoni A, Martelli PL, Fariselli P, Casadio R: **BaCelLo: a balanced subcellular localization predictor**. *Bioinformatics* 2006, **22**:e408-416.

14. Nair R, Rost B: **Mimicking cellular sorting improves prediction of subcellular localization**. *J Mol Biol* 2005, **348**:85-100.

15. Maurer-Stroh S, Eisenhaber F: **Refinement and prediction of protein prenylation motifs**. *Genome Biol* 2005, **6**:R55.

16. Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X: **CSS-Palm 2.0: an updated software for palmitoylation sites prediction**. *Protein Eng Des Sel* 2008, **21**:639-644.

17. Julenius K: **NetCGlyc 1.0: prediction of mammalian C-mannosylation sites**. *Glycobiology* 2007, **17**:868-876.

18. Julenius K, Molgaard A, Gupta R, Brunak S: **Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites**. *Glycobiology* 2005, **15**:153-164.

19. Gupta R, Brunak S: **Prediction of glycosylation across the human proteome and the correlation to protein function**. *Pac Symp Biocomput* 2002:310-322.

20. Blom N, Gammeltoft S, Brunak S: **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites**. *J Mol Biol* 1999, **294**:1351-1362.

21. Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM: **Identification, analysis, and prediction of protein ubiquitination sites**. *Proteins* 2010, **78**:365-380.

22. Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y: **Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0**. *Proteomics* 2009, **9**:3409-3412.

23. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server**. *Bioinformatics* 2000, **16**:404-405.

24. Cole C, Barber JD, Barton GJ: **The Jpred 3 secondary structure prediction server**. *Nucleic Acids Res* 2008, **36**:W197-201.

25. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C: **A generic method for assignment of reliability scores applied to solvent accessibility predictions**. *BMC Struct Biol* 2009, **9**:51.

26. Capriotti E, Fariselli P, Casadio R: **I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure**. *Nucleic Acids Res* 2005, **33**:W306-310.

27. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY: **FoldUnfold: web server for the prediction of disordered regions in protein chain**. *Bioinformatics* 2006, **22**:2948-2949.

28. Dosztanyi Z, Csizmok V, Tompa P, Simon I: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**. *Bioinformatics* 2005, **21**:3433-3434.

29. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc* 2009, **4**:1073-1081.

30. Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:166.

31. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**:3894-3900.

32. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P: **PROSITE: a documented database using patterns and profiles as motif descriptors**. *Brief Bioinform* 2002, **3**:265-274.

33. • Gould CM, Diella F, Via A, Puntervoll P, Gemund C, Chabanis-Davidson S, Michael S, Sayadi A, Bryne JC, Chica C *et al.*: **ELM: the status of the 2010 eukaryotic linear motif resource**. *Nucleic Acids Res* 2010, **38**:D167-D180.
The ELM ressource covers a large number of linear motifs and presents them in an appealing format. As such it is an excellent example of a centralised prediction pipeline.

34. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A *et al.*: **ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins**. *Nucleic Acids Res* 2003, **31**:3625-3630.

35. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2010, **38**:D211-222.

36. Jimenez RC, Quinn AF, Garcia A, Labarga A, O'Neill K, Martinez F, Salazar GA, Hermjakob H: **Dasty2, an Ajax protein DAS client**. *Bioinformatics* 2008, **24**:2119-2121.

37. Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P *et al.*: **Integrating biological data — the Distributed Annotation System**. *BMC Bioinformatics* 2008, **9(Suppl 8)**:S3.

38. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, Hostetter G, Boguslawski S, Moses TY, Savage S *et al.*: **A transforming mutation in the pleckstrin homology domain of AKT1 in cancer**. *Nature* 2007, **448**:439-444.

39. • Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, Yaffe MB, Pawson T: **NetworKIN: a resource for exploring cellular phosphorylation networks**. *Nucleic Acids Res* 2008, **36**:D695-699.
This paper describes a resource for looking at protein kinases in relation to their substrates and corresponding phosphoprotein-binding domains on a large scale.

40. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S: **A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes**. *Proc Natl Acad Sci U S A* 2008, **105**:20870-20875.

41. •• Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N *et al.*: **A human phenome–interactome network of protein complexes implicated in genetic disorders**. *Nat Biotechnol* 2007, **25**:309-316.
This paper demostrates how to integrate quality-controlled interactions of human proteins with a validated, computationally derived phenotype similarity score, thus permitting identification of previously unknown complexes likely to be associated with disease.

42. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2010, **38**:D355-360.

43. Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list**. *Nucleic Acids Res* 2006, **34**:D504-506.

44. Xu C, Goodz S, Sellers EM, Tyndale RF: **CYP2A6 genetic variation and potential consequences**. *Adv Drug Deliv Rev* 2002, **54**:1245-1256.

45. Daniels G: **The molecular genetics of blood group polymorphism**. *Transpl Immunol* 2005, **14**:143-153.

46. Sturm RA: **Molecular genetics of human pigmentation diversity**. *Hum Mol Genet* 2009, **18**:R9-17.

47. Steinberg MH: **Genetic etiologies for phenotypic diversity in sickle cell anemia**. *ScientificWorldJournal* 2009, **9**:46-67.

48. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A: **Adaptations to climate in candidate genes for common metabolic disorders**. *PLoS Genet* 2008, **4**:e32.

49. Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC: **Effects of purifying and adaptive selection on regional variation in human mtDNA**. *Science* 2004, **303**:223-226.

50. •• Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R *et al.*: **Ancient human genome sequence of an extinct Palaeo-Eskimo**. *Nature* 2010, **463**:757-762.
This paper describes the sequencing and genome-to-phenome studies of an ancient human genome.

51. •• Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK *et al.*: **Sequencing and analysis of Neanderthal genomic DNA**. *Science* 2006, **314**:1113-1118.
Efforts to reconstruct Neanderthal characteristics from SNPs were portrayed as the first Neandarthal model face in the October 2008 issue of National Geographic. This paper as well as the Palaeo-Eskimo genome sequence paper attempt to reconstruct characteristics of truly anonymous individuals, and many of such phenotypes are based on protein non-synonymous coding changes.

52. Voth DE, Heinzen RA: **Sustained activation of Akt and Erk1/2 is required for *Coxiella burnetii* antiapoptotic activity**. *Infect Immun* 2009, **77**:205-213.

Chapter 5

Variant calling with NGS

NGS technologies rely on randomly fragmenting the DNA into small segments and produce millions of short sequencing reads (25-500 bp) coming from those segments. After several rounds of this fragmentation and sequencing, multiple overlapping reads are obtained which can be either mapped back to the reference genome or assembled into a continuous sequence. Variant calling is then simply identifying the positions where the sequenced base is different from the base in the reference genome. This task is complicated by many factors including sequencing errors, alignment errors and low coverage. This chapter describes the pipeline used for NGS data handling for SNP calling as well as the common obstacles encountered during analysis.

5.1 Raw read quality control

Every base in each sequencing read is assigned a quality score derived from the noise estimates of the image generated by a sequencing platform. The standard for reporting per-base quality scores is the Phred quality score [41] defined as:

$$Q_{Phred} = -10log_{10}P(error)$$

This means that a Phred score of 20 corresponds to 1 in 100 probability that a given base is called incorrectly. The base quality scores are used for quality control of the raw reads and further during alignment steps and downstream analysis. Typically reads with average low qualities are discarded from further analysis and low quality ends of reads are trimmed off to

prevent alignment problems. Another important step is removal of sequencing adaptor sequences which might have accidentally been sequenced and could bias the alignment. It is also a good practice to investigate the overall quality of raw data by checking for any deviations of per-base sequence quality and GC content, as well as overrepresented sequences and k-mers. Those few simple steps can improve the overall read quality dramatically and prevent from including reads coming from contamination by sequence adaptors or of bacterial or viral origin in the downstream analysis.

5.2 Alignment

After filtering of raw reads, the high-quality reads are mapped back to the reference genome. This task is complicated by many possible mismatches in the reads caused either by true variants or indels or by sequencing errors. With the short read length it is particularly difficult to correctly align the sequencing reads in highly repetitive regions and regions with high levels of diversity, such as the major histocompatibility complex (MHC) region. This can be partly overcome by using longer reads and paired-end sequencing allowing for more accurate estimation of the origin of the read.
The existing widely used alignment algorithms are mostly based either on hashing of the reads or the reference genome (e.g. Novoalign [http://novocraft.com/], Stampy [76]), or on the Burrows-Wheeler transform (BWT) algorithm for effective data compression (e.g. BWA [73], Bowtie [71]). Detailed description of the algorithms is beyond the scope of this thesis, but in general the hash-based methods produce more accurate results, while the BWT-based aligners are considerably faster and more memory-efficient. In analysis presented in Paper II, III and IV the chosen aligner was BWA, which seemed to be the gold-standard tool at the time of analysis.
The need to perform the mapping in a fast and efficient manner may sometimes compromise the accuracy of mapping. Since the correct alignment is crucial for accurate SNP calling, it is important to perform a few alignment refining steps. Presence of indels often affects the alignment around them and might lead to discovery of many false-positive variations, therefore performing additional local realignment around indels using Smith-Waterman algorithm can significantly improve the alignment accuracy [37]. Most of the available sequence alignment programs report a mapping quality score for each aligned sequence, which quantifies the probability that the given alignment is correct. Therefore usually one of the first steps in evaluating of the quality of the alignment would be discarding the reads with low mapping scores.
Another common source of SNP calling errors may arise from excess of PCR duplicates, which may lead to multiple counting of an allele arising in fact from the same source biological sequence. Thus in order to only account for unbiased variant allele evidence, it is important to remove any sequences

with the same sequence, start site and orientation which could suggest that they represent multiple reads of the same unique DNA fragment amplified by PCR.
Finally, variant calling takes into account the quality scores of the appropriate bases, therefore it is important that the scores are well-calibrated. It was shown that the quality scores produced by base-calling algorithms do not always accurately reflect the true error rates [16]. The Genome Analysis Toolkit [80] implements base-quality recalibration by comparing the sequenced sample to the reference genome at sites with no known SNPs at the same time accounting for the machine cycle and dinucleotide content.

5.3 SNP calling

SNP calling (or: variant calling) determines at which positions there are polymorphisms, while genotype calling is defining the exact genotype for each individual at those positions. The standard approach to SNP calling is to compare the proportion of high quality sequences containing the reference and the non-reference allele at a given position. In principle, the same method is used for detection of short indels and multiple nucleotide polymorphisms (MNPs) by SNP calling programs, with the only difference that gapped mapping technique is used to identify locations where there is either a compression or an expansion of the genomic sequence. Conventionally, if the fraction of reads with the alternative allele is between 20 - 80% then a heterozygous genotype is called, and a homozygous genotype is called otherwise [90]. While this works well for high coverage data (>20x), for moderate sequencing depths applying probabilistic methods which provide a posterior probability for each genotype accounts for the uncertainty in genotyping. SNP and genotyping calling procedures can be further improved by incorporating prior information about the known polymorphic sites, as well as the LD patterns between sites. The resulting list of polymorphisms can be further refined by examining sequencing depth at the variant site, distribution of base quality scores, as well as biases in strand representation and position along the read. Additionally, clusters of SNPs in a small region often indicate errors in alignment arising for example due to highly similar paralogous sequences, therefore those should be carefully examined and possibly filtered out.
Identification of novel rare variants or point mutation requires much more stringent filtering than genotype calling at known sites. In the latter case prior knowledge about observed alleles and their expected frequencies can increase our confidence in the observed genotype. For variant detection purposes one has to ensure sufficient sequencing depth, as well as closely investigate all possible sources of sequencing, alignment and SNP calling errors. Identifying mutations in tumour samples can be particularly challenging as cancer DNA commonly acquires copy number alterations resulting in different number of copies of whole chromosomes or its parts, and therefore

the assumption of diploid genome held by most widely used variant callers might not be best suited for variant calling in these samples. Moreover, due to contamination with normal tissue and the heterogeneity of the tumour, identifying mutations in cancer DNA requires very high sequencing depth to obtain reliable results.
SAMtools [74] SNP caller was used in analyses presented in Paper II, III and IV and at the time of analysis this tool seemed to be the most commonly used software for this task. Recently, also Unified Genotyper from Genome Analysis Toolkit [80] receives a lot of attention, using similar method for SNP detection with additional features including applying prior information on known SNP sites. There are many other variant callers available based on different assumptions than the two aforementioned tools, including FreeBayes [44] which includes haplotype information up to the length of sequencing reads, the Complete Genomics caller [20] using local assembly or Cortex [60] performing almost complete *de novo* assembly using coloured de Bruijn graphs for variant detection. Due to greater complexity of these tools, the computational power required to run them is much bigger than for the traditional variant callers. The performance of the new tools will have to be comprehensively reviewed and assessed on an independent dataset before they are routinely implemented in NGS pipelines.

The steps commonly used in handling NGS data for variant detection are summarized in Figure 5.1. Additional steps are included specific for multiplexed targeted sequencing technology used in Papers II, III and IV in this thesis: sorting of the raw reads by the unique barcode identifying the sample of origin and filtering out SNPs residing outside of the targeted regions.

5.4 CNV calling

Traditionally CNVs are detected by applying a circular binary segmentation (CBS) algorithm [93] on the signal intensities from individual SNP markers on SNP arrays. The algorithm divides a genomic region into segments and tests whether the normalized signal intensities of the neighbouring segments are different, suggesting that they represent different copy number states. After identifying the segments of the same copy number, the segment ends are joined forming a circle to perform another likelihood ratio test whether the signal intensities have different means. An alternative to this method GADA (Genome Alteration Detection Algorithm) [98] was applied to detect CNVs from the Affymetrix Gene-Chip Mapping 500K array set for 62 childhood ALL samples and 200 healthy controls in Paper V in this thesis. Briefly, GADA algorithm starts with developing a compact linear algebra representation for the genome copy number from normalized probe intensities, then applies sparse Bayesian learning to infer copy number changes locations, and finally a backward elimination procedure to rank the inferred breakpoints is applied.

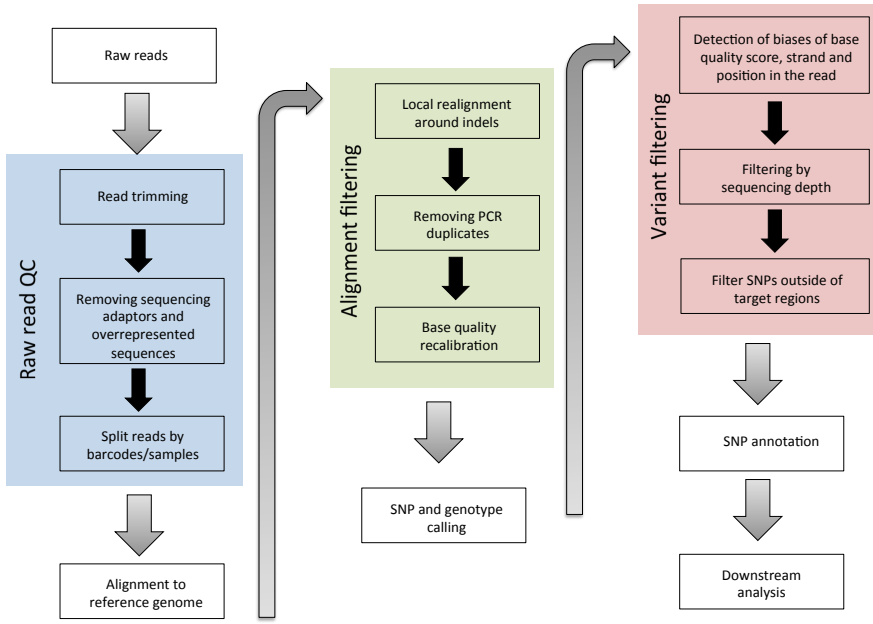


Figure 5.1. The pipeline used to call SNPs and genotypes from raw sequencing reads obtained from multiplexed targeted sequencing experiment.

NGS data can also be used to investigate copy number variations in genomic samples with a much better resolution than SNP arrays. The best suited for this purpose is whole genome sequencing where the mean coverage of the genome can serve as a baseline value and any significant deviations from this coverage can indicate copy gain or loss. When using NGS data to infer copy number it is important to take into account the GC content of the region, as for instance on the Illumina platform regions with extremely low (<20%) or high (>60%) GC content are known to be underrepresented among the sequencing reads [12]. Copy number variations can also be estimated from targeted sequencing data when data for a large population of samples generated in comparable way is available. In Paper II in this thesis we have estimated the copy numbers of the high frequency deletions of *GSTT1* and *GSTM1* genes. To estimate the copy number, a depth ratio was calculated from the number of reads in the targeted genomic region normalized by size of the region and total number of reads for the sample (Figure 2, Paper II). The GSTs copy numbers estimated by this method were in 100% concordant

with the copy number states assayed by multiplexing PCR for the 42 samples examined by both methods. This method for estimating deletion states of GST genes has also been applied in two other papers not included in this thesis: Borst *et al.* 2012 [14] and Edsgärd *et al.* 2012 (manuscript accepted with minor revisions by Frontiers in Cancer Endocrinology).

5.5 Other challenges

As the cost of sequencing declines, the number of NGS applications grows, however NGS data analysis still presents substantial bioinformatics challenges apart from the ones already mentioned in this chapter. To start with, the amount of data produced by sequencers is overwhelming and even compressed it requires massive disk storage space. In the data analysis pipeline multiple intermediate data files are generated which also need to be temporarily stored, and since those files are generally large - the softwares processing them require a lot of CPU power, time and memory. The available computational infrastructure might be not sufficient for smaller labs to conduct NGS analysis on bigger scale. However this obstacle can be easily overcome by cloud computing services, which lease computational power to those in need. Another complication is the diversity of the sequencing platforms available and lack of common standards for reporting base calling qualities despite the most commonly used FASTQ format for storing biological sequences. Even different sequencing platforms from the same manufacturer can have different base quality representations, which can cause a lot of confusion for an inexperienced user. Furthermore, different sequencing platforms require different methods for analysis, for instance when analysing Illumina data one has to take into account that base quality decreases along the read, while when analysing Roche data one has to be aware of frequent errors around homopolymer regions. Data produced by SOLiD sequencers are encoded in colour space with each colour representing two consecutive bases and therefore the SOLiD data requires its own data format and specific alignment programs making use of this information. With many new sequencing platforms coming to the market we can expect even more data formats to come and new pitfalls of sequencers to take care of. The choice of platform is followed by vast choice of software to perform the analysis, however without gold standards established in the field and lack of comprehensive review articles it is difficult to know which software is appropriate for the task. Moreover, the programs are often dependent on specific data format produced by another upstream software, which complicates freely combining the individual components of a pipeline.

Chapter 6

Hypothesis-driven SNP selection and assay

This chapter describes the motivation for the genome-wide selection of genomic variation with potential clinical importance for childhood ALL and development of custom genotyping method by means of multiplexed targeted sequencing applied in Papers II, III and IV in this thesis.

6.1 SNP selection

In order to fully investigate the pharmacogenomics in childhood acute lymphoblastic leukaemia there is a need to go beyond the limited candidate gene approaches and explore all potentially clinically important variations. For this purpose an extensive literature curation was conducted collecting the current state of art knowledge of drug pharmacokinetics and pharmacodynamics in ALL, disease mechanisms and response to chemotherapy. This included previously published functional studies of polymorphisms affecting metabolism, transport, targets, as well as drug-related toxicity of the 13 most commonly administered chemotherapeutic drugs [33]. Mechanisms of action and the most important pharmacogenetic domains of the drugs are descibed in Chapter 3.2. The list of genes and their respective polymorphisms was further expanded to include additional aspects of response to chemotherapy, including immune system functions, apoptosis, mitosis, DNA repair mechanisms, drug absorption, metabolism, excretion and cellular transport, as well as drug targets and metabolic pathways affected by the drugs. Information about known drug targets and drug interactions was collected from the Pharmacogenomics Knowledge Base [54], DrugBank [131] and Comparative

Toxicogenomics Database [35] (all version: 2008). Further, high confidence human interactome data [68] was used to expand the list of relevant molecules with the first order protein-protein interaction partners. All known polymorphic sites were extracted within the genomic boundaries of the resulting 969 genes using Ensembl API version 57 [58] (based on dbSNP v.130 [118]), and then filtered based on their consequence to the transcript predicted with Ensembl Variant Effect Predictor [81]. Only variations with likely functional effect on the resulting protein sequence or its expression were selected, including variations with following effects on their transcripts:

- Non-synonymous coding
- Frame-shift coding
- Stop gained
- Stop lost
- Splice site
- Regulatory region
- Within mature miRNA
- Within non-coding gene

Additionally, information about polymorphic miRNA target sites within 3' UTR regions of the selected genes was retrieved from Patrocles database [55].

The full list of resources used in the SNP selection process is summarized in Table 6.1. The SNP selection process is illustrated schematically in Figure 1 of Paper II in this thesis.

6.2 Available genotyping methods review

After careful selection of the panel of potentially clinically relevant SNPs followed the choice of appropriate method to assay the selected polymorphisms. Commercially available genotyping arrays commonly used for GWAS were not suited for this application as the SNP arrays are designed to test random genome-wide variations with high minor allele frequencies. Majority of the assayed variations reside in the non-coding regions of the genome, and therefore the overlap between the list of SNPs selected for genotyping in this study and the SNPs assayed with commercially available platforms was little. In addition, studies conducted using GWAS principles are not hypothesis-driven, and due to the requirements for strict statistical corrections - thousands of samples are needed to achieve genome-wide significance

Resource	URL	Description
Pharmacogenomics Knowledge Base (PharmGKB) [54]	www.pharmgkb.org	Curated knowledge about the impact of genetic variation on drug response including dosing guidelines and drug labels, potentially clinically actionable gene-drug associations and genotype-phenotype relationships. The database comprises variant annotations curated for more than 700 genes in relation to pharmacogenetic of more than 400 drugs and 92 pathways describing drug pharmacokinetics or pharmacodynamics.
DrugBank [131]	www.drugbank.ca	Database of drug targets, contains annotations for 6,711 drug entries linked to 4,227 non-redundant proteins acting as drug targets, enzymes, transporters or carriers.
Comparative Toxicogenomics Database (CTD) [35]	ctdbase.org	Database collecting more than 18 million toxicogenomic relationships, including associations between genes, chemicals and diseases.
InWeb [68]	*in-house database*	Meta-database of protein-protein interactions comprising several other resources: BIND [8], BioGRID [121], CORUM [111], DIP [113], IntAct [53], HPRD [97], MINT [22], MPact [49], MPPI [96] and OPHID [18].
Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database [92]	www.genome.jp/kegg/pathway.html	Database comprising a large collection of biological pathways annotated through literature-based searches.
Reactome [31]	www.reactome.org	Database comprising 1,326 manually curated and peer-revied biological pathways.
Ensembl	www.ensembl.org	A centralized resource for studying genome sequences, genome variations, regulation and comparative genomics. Variation database collects data from dbSNP [118] together with annotations on the effect of the SNP on the transcript and on associated phenotypes from HGMD [28], COSMIC [42], NHGRI GWAS catalog [56], OMIM [50], UniProt [9] and others.
Patrocles [55]	www.patrocles.org	Database with annotations of polymorphisms within microRNAs and their targets.

Table 6.1. This table described the publicly available resources used for the SNP selection for Papers II, III and IV.

of the results. Designing of a custom SNP array was not feasible either due to high minimum order quantity requirements much larger than the number of patients included in this study and therefore setting the costs of this method too high for implementation in clinical settings.

Target enrichment [46] methods for next-generation sequencing (Figure 6.1) are widely used for exome sequencing, GWAS follow-up and resequencing studies. In a similar manner, capture baits can be custom-designed complementary to regions harbouring the SNPs of interest from this study. This method has the advantage of genotyping not only at the exact position of the target SNPs, but also provides the opportunity to detect any possible variation in their vicinity. The capture baits are 120 nucleotides long and are therefore more specific than the microarray probes, and due to the length of the baits it is also possible to detect short indels. Another important advantage of this method is flexibility of the design allowing for modifying the bait sequences with every order based on observed performance.

6.3 Multiplexing - pilot study

The costs of targeted sequencing, even though much lower than whole genome sequencing costs, are still too high to apply in clinical settings to screen hundreds of patients. The major cost of the experiments is the cost of the target capture kit and the cost of sequencing itself. The approximate size of the total region harbouring all the variations with potential clinical significance for childhood ALL study is 3.5 mega base pairs. With current NGS technologies the output of the sequencer is sufficient to sequence this region to a very high depth. For the purpose of genotype calling at known SNP loci we estimated a sequencing depth of 10x to be sufficient to produce reliable results. By labelling individual samples with short barcode sequences it is possible to sequence several samples in one sequencing lane and identify the origin of the sequencing reads through the unique barcode. By multiplexing the samples the whole capacity of the sequencer output can be used to produce satisfying results for multiple samples in one experiment. Multiplexing samples for sequencing has been shown to be successful [91] and to significantly decrease the cost per sample. In an attempt to decrease the cost even further, we have tested whether multiplexing of the sample before the target capture step is possible and whether it produces reliable results and the desired sequencing depth per sample.
The aim of the pilot study presented in Paper II was to assess applicability of multiplexed targeted sequencing for genotyping and to investiagate the influence of the study design on the achieved performance. We demonstrate the positioning of the bait with respect to targeted variation can affect the results and that using two overlapping capture baits instead of one yields higher sequencing depth at the position of variation, as well as more uniform coverage in the region surrounding it (Figure 6.2).

Further we examined the physico-chemical properties of the individual baits and compared the achieved sequencing depth at the regions targeted by those baits (Figure 6.3). We observed that especially extreme GC content and low complexity of the targeted region influenced the results. Additionally, cross-hybridization of the baits and their potential for self-folding also had an

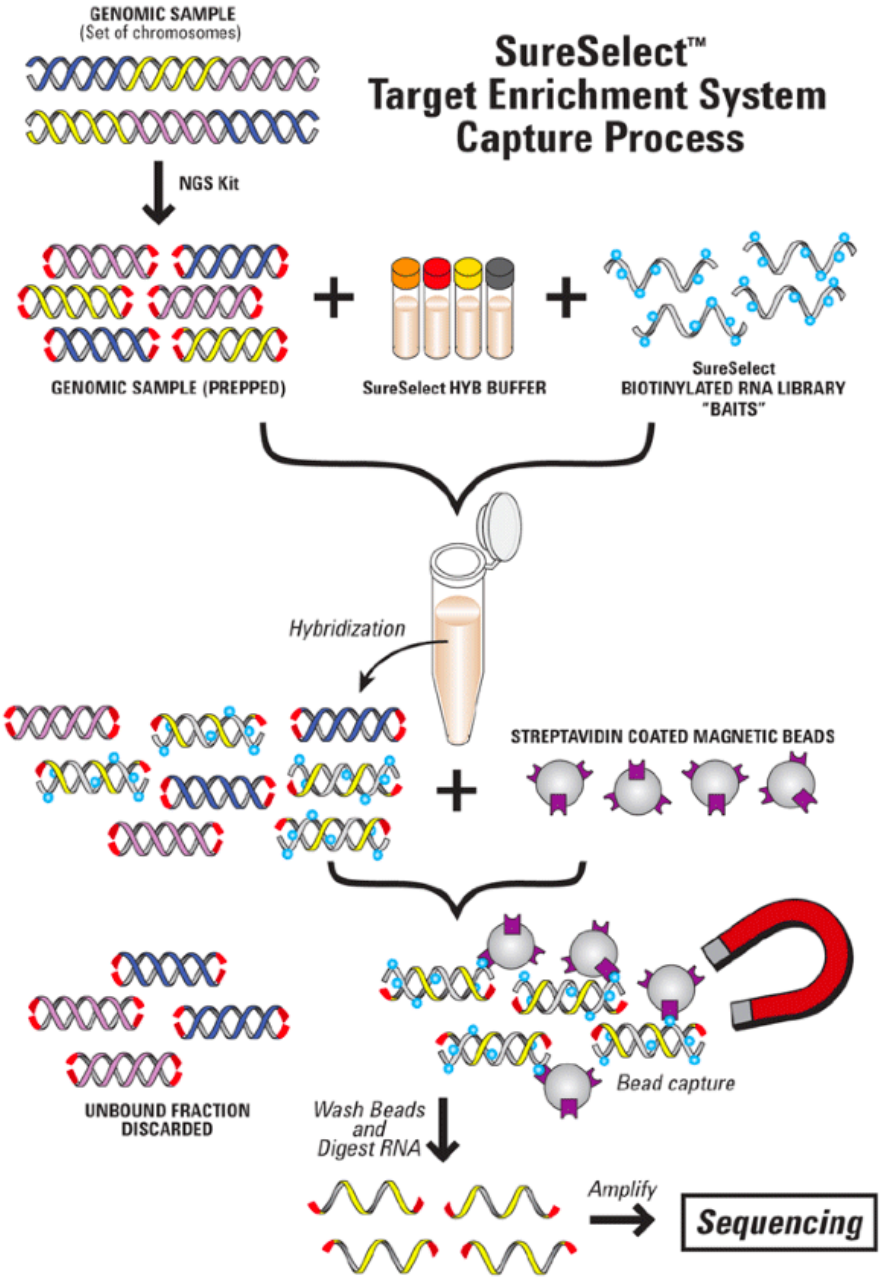


Figure 6.1. The workflow of Agilent's Sure Select Target Enrichment System. The genomic sample is fragmented into shorter fragments and hybridized together with biotinylated library of RNA capture baits. Hybridized RNA-DNA duplexes are pulled down with streptavidin coated magnetic beads. After washing the beads, the RNA baits are digested and the subset of initial DNA is ready for sequencing. Source: `www.genomics.agilent.com`.

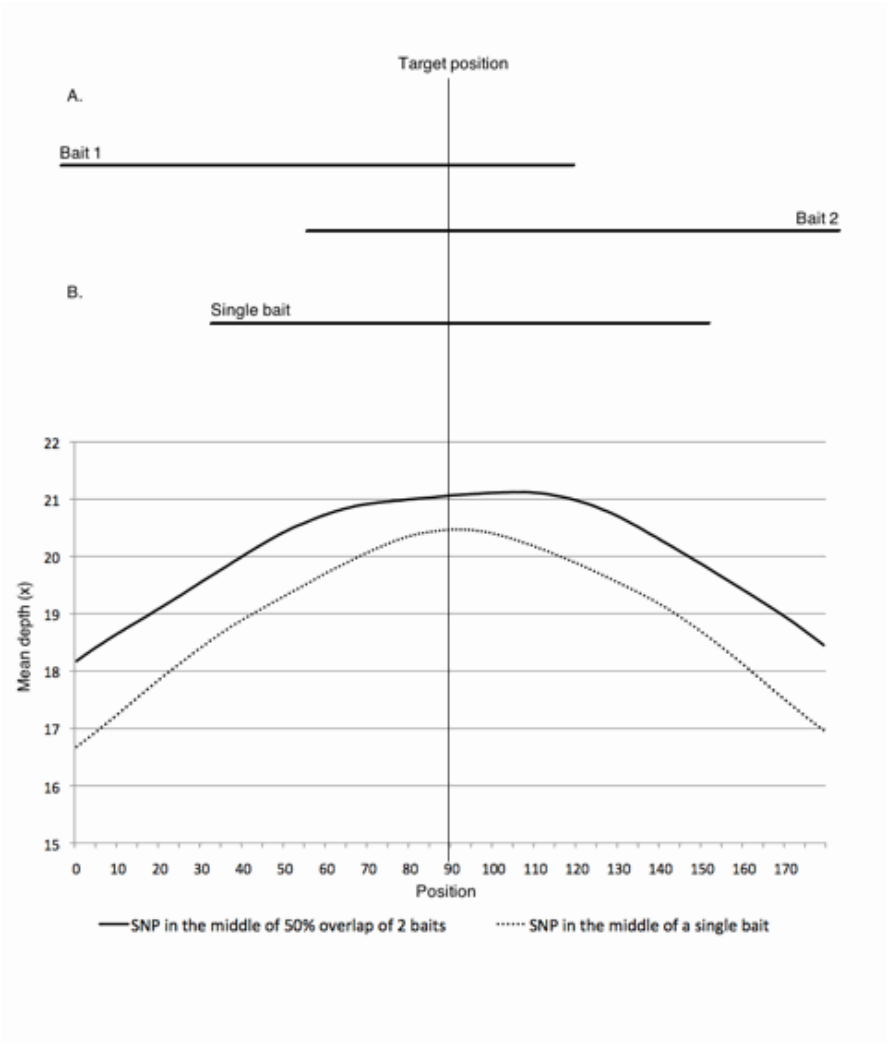


Figure 6.2. Average depth coverage at the target positions and their 90 nucleotide flanking sequence. The length of the flanking sequence corresponds to the length of two 120 nucleotide long consecutive baits overlapping with 50% of their length. A. The target positions were defined as the base pair positioned exactly in the middle of the 50% overlap of the two consecutive baits. This location of the target positions with respect to the baits was chosen for the performance analysis due to more uniform distribution of the coverage depth in the immediate surrounding of the target position. B. The target positions were defined as the base pair exactly in the middle of each single bait. The depth coverage drops in the proximity of the target position and in general slightly lower depth coverage is achieved. Figure is a part of Supplementary Material for Paper II.

impact on the capture experiment efficiency. Examination of targeted regions which did not get any sequencing coverage revealed that 97% of the baits targeting these regions were prone to encounter cross-hybridization problems, and 56% of them had a high probability of self-folding. This demonstrates that designing the baits more carefully can increase the achieved mean coverage depth.

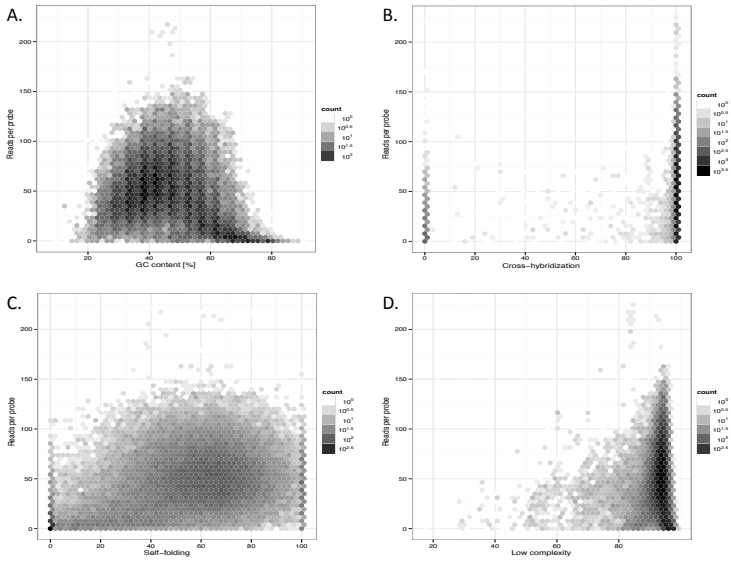


Figure 6.3. Performance of the baits depends on their physical and cross-hybridizational properties. For all the baits used in the experiment here is shown A. the dependence on the GC content of the bait, B. the potential for cross-hybridization or C. self-folding of the bait and D. presence of low-complexity regions. The values in B, C and D are the scores obtained from the OligoWiz probe design software, where 0 denotes a potentially problematic bait and 100 a good bait with respect to the investigated parameter. Figure is a part of Supplementary Material for Paper II.

Finally, we examined the combined influence of the amount of samples pooled in one sequencing lane together with the properties of the used baits (Figure 6.4). Pooling of 4, 6 and 12 samples was compared to an experiment using a single sample by assessing the amount of targeted genomic variations sequenced at different thresholds of minim required requencing depth. Even though we observed that the performance slightly decreased with an increasing number of pooled samples, the distribution of reads was well balanced

even with 12 pooled samples and a mean depth of at least 10x was achieved for approximately 88% of the target positions. The amount of pooling can be optimized depending on the desired target size and need for high coverage and the resulting sequencing depth can be improved by more methodical custom bait design to achieve more efficient capture .

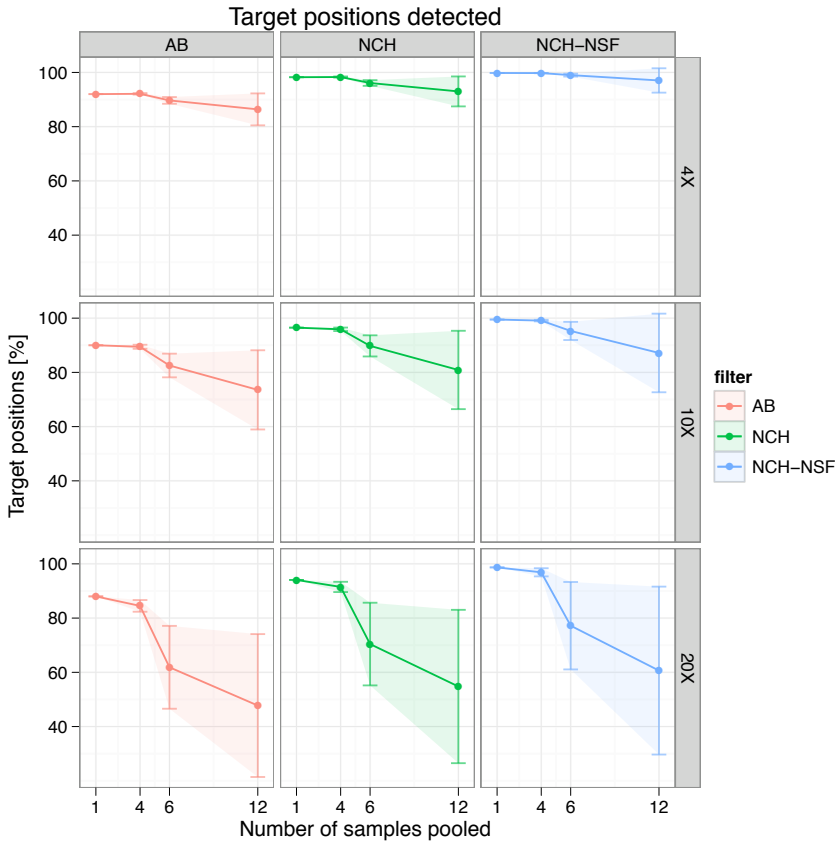


Figure 6.4. Percentage of the target regions achieving a minimum sequencing depth threshold of 4x, 10x and 20x. The target regions are defined as contigs of consecutive baits for the set of all SureSelect Human X Chromosome Demo Kit baits (AB), the reduced set of non cross-hybridizing baits (NCH) and the reduced set of non cross-hybridizing and non self-folding baits (NCH-NSF). The error lines represent the standard deviation. Figure is a part of Supplementary Material for Paper II.

6.4 Paper II - Multiplexing before capture

In the following paper a novel cost-effective strategy of pooling of samples before capture is described. We have compared target capture of a single sample with capture of pools of 4, 6 and 12 samples. The results show that while the distribution of reads between samples is not entirely balanced, the multiplexing of samples works and the SNP calls obtained from this method achieve high concordance with standard SNP array calls. We also demonstrate that detection of copy number variations based on sequencing depth is possible with this data and produces accurate results. The issue of optimal bait design is also addressed in the article. We demonstrate that baits with not optimal physico-chemical properties perform worse in the experiment. We observed that regions with extreme GC content or low complexity are difficult to target, and we show that baits with high likelihood of cross-hybridization and self-folding might cause further performance issues. Finally, we demonstrate that mitochondrial DNA is generally overrepresented in genomic samples and, even with targeted sequencing approach, reads from mitochondrial regions can dominate the sequencer output if some capture baits can hybridize to those regions. Therefore we recommend to remove any such baits from the capture library design to maximize the "on target" sequencing coverage.
Supplementary Information accompanies the paper on the Leukemia website (`http://www.nature.com/leu`) and parts of the material are included as figures in the preceding chapter.

Leukemia (2011), 1–6

www.nature.com/leu

ORIGINAL ARTICLE

Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia

A Wesolowska[1,8], MD Dalgaard[2,8], L Borst[3,8], L Gautier[1], M Bak[4], N Weinhold[1], BF Nielsen[2], LR Helt[3], K Audouze[1], J Nersting[3], N Tommerup[4], S Brunak[1,6], T Sicheritz-Ponten[1,7], H Leffers[2], K Schmiegelow[3,5] and R Gupta[1]

[1]*Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Kgs. Lyngby, Denmark;* [2]*Department of Growth and Reproduction, University Hospital Rigshospitalet, Copenhagen, Denmark;* [3]*Department of Pediatrics, Pediatric Clinic II, Juliane Marie Centre, University Hospital Rigshospitalet, Copenhagen, Denmark;* [4]*Department of Cellular and Molecular Medicine, Wilhelm Johannsen Centre for Functional Genome Research, Panum Institute, Copenhagen, Denmark;* [5]*Faculty of Medicine, Institute of Gynecology, Obstetrics and Pediatrics, University of Copenhagen, Copenhagen, Denmark;* [6]*Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark and* [7]*Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby, Denmark*

Genetic variants, including single-nucleotide polymorphisms (SNPs), are key determiners of interindividual differences in treatment efficacy and toxicity in childhood acute lymphoblastic leukemia (ALL). Although up to 13 chemotherapeutic agents are used in the treatment of this cancer, it remains a model disease for exploring the impact of genetic variation due to well-characterized cytogenetics, drug response pathways and precise monitoring of minimal residual disease. Here, we have selected clinically relevant genes and SNPs through literature screening, and on the basis of associations with key pathways, protein–protein interactions or downstream partners that have a role in drug disposition and treatment efficacy in childhood ALL. This allows exploration of pathways, where one of several genetic variants may lead to similar clinical phenotypes through related molecular mechanisms. We have designed a cost-effective, high-throughput capture assay of ~25 000 clinically relevant SNPs, and demonstrated that multiple samples can be tagged and pooled before genome capture in targeted enrichment with a sufficient sequencing depth for genotyping. This multiplexed, targeted sequencing method allows exploration of the impact of pharmacogenetics on efficacy and toxicity in childhood ALL treatment, which will be of importance for personalized chemotherapy.
Leukemia advance online publication, 18 March 2011; doi:10.1038/leu.2011.32
Keywords: multiplexed genotyping; next-generation sequencing; target-enrichment; clinically relevant SNPs; childhood acute lymphoblastic leukemia

Introduction

The cure rate for childhood acute lymphoblastic leukemia (ALL) after first-line therapy approaches 80–85%.[1] To further improve prognosis, extensive personalization of therapy based on extensive targeted genetic analyses will be required.[2] The overall goal is to avoid unnecessary adverse and life-threatening toxicities, unacceptable late effects due to overtreatment and risk of relapse due to undertreatment.[3,4] Several studies have indicated that for patients with the most favorable genetic variants, event-free survival may be more than 90%.[5,6] High-throughput technologies in combination with imputing allow genome-wide mapping of genetic variants that subsequently can be associated with treatment failures or specific toxicities.[7] In addition to the genes and single-nucleotide polymorphisms (SNPs) already known to be involved in drug disposition or specific toxicities (for example, thrombosis, osteopenia or immune function), such genome-wide variation studies (GWAS, genome-wide association study) are likely to reveal important variations in genes not previously linked to the biological issue in question. However, extensive research including clinical trials will subsequently be needed before this new information can be implemented in childhood ALL treatment protocols. Furthermore, the impact of the genetic variations can often only be fully understood within the frame of a specific treatment protocol.[2] Finally, current commercially available solutions for GWAS for hypothesis-driven genetic investigations are not easily applied clinically, as the techniques and commercial platforms are either designed to explore random variations across the genome that rarely cover all variations of interest for a specific study, or the costs of custom-made approaches are too high for implementation in clinical settings. We here describe a novel multiplexing method enabling us to screen childhood ALL patients for ~25 000 clinically relevant SNPs simultaneously, targeted by custom-designed baits. Furthermore, eight childhood ALL samples are pooled together before capture enrichment, making this a very cost-effective platform, allowing future targeted genetic mapping of large cohorts of patients. The choice of pooling 8 patient samples was based on results from a pilot study, where we pooled and sequenced 4, 6 and 12 test samples, respectively, labeled with different barcodes.

Correspondence: Dr R Gupta, Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, Building 208,Kgs. Lyngby 2800, Denmark.
E-mail: ramneek@cbs.dtu.dk
[8]Joint first authorship.
Received 21 December 2010; accepted 11 January 2011

Materials and methods

Samples

A total of 48 samples from Danish childhood ALL patients (aged between 1 and 15 years at the time of diagnosis) diagnosed with B-cell precursor or T-lineage ALL and enrolled in the Nordic Society for Pediatric Hematology and Oncology (NOPHO) ALL-2000 protocol were included. For the pilot study, four human DNA and two HapMap DNA test samples were used in different combinations (Supplementary Table 1). The study was approved by The Danish Data Protection Agency (2007-41-1289) and The Committee on Biomedical Research Ethics (H-D-2007-0100, KF 01 265848).

Library preparation, pooling, target enrichment and sequencing

DNA shearing and library preparations were performed according to the SureSelect Target Enrichment System protocol version

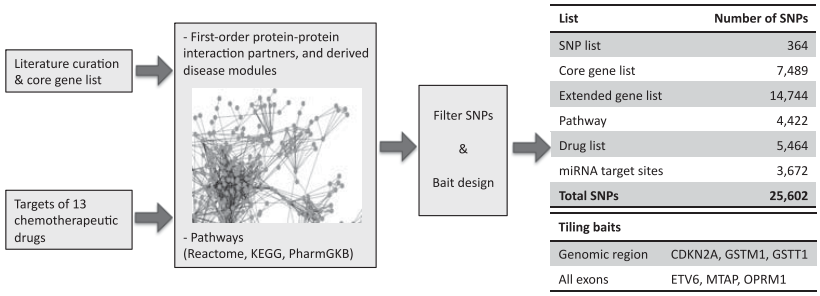
1.2 April 2009 (Agilent Technologies, Santa Clara, CA, USA) with minor modifications. Briefly, 3 μg of genomic DNA was sheared by Covaris S2 System (Covaris Inc., Woburn, MA, USA) using 10% duty cycle, intensity of 5, cycles per burst of 200 for 6 cycles of 60 s, following purification of the DNA fragments by QIAquick PCR purification spin columns (Qiagen, Hilden, Germany). After each reaction, a purification step was performed. Then end repair was performed (by applying T4 DNA polymerase, T4 phosphonucleotide kinase and Klenow fragment enzyme) and 3′ end A-overhangs were produced (by applying Klenow 3′–5′exo minus). Custom-made adapters containing unique barcodes of four bases each were prepared. The complementary oligos (Supplementary Table 2; DNA technology A/S, Risskov, Denmark) were dissolved in DNase-free water to a final concentration of 300 μM. Complementary oligonucleotide pairs were mixed in ratio 1:1 in 1 × annealing buffer (10 × buffer contained 100 mM Tris-HCL, pH 8.1; 0.5 M NaCl). The barcoded adapter mix was heated to 90 °C for 2 min, then cooled down to 30 °C at a rate of 2 °C per minute and diluted to a working concentration of 15 μM. After ligation of the adapters to the DNA fragments, the fragments were size selected in the range of 150–250 bp by 4% agarose gel electrophoresis and excised. The DNA libraries were amplified applying Phusion High-Fidelity PCR Master Mix (Finnzymes, Espoo, Finland) with a denaturation time of 30 s at 98 °C, followed by 14 cycles of denaturation at 98 °C for 10 s, annealing at 65 °C for 30 s and extension at 72 °C for 30 s. Final extension was performed at 72 °C for 5 min. DNA quantity and quality was checked on a NanoDrop ND- 1000 UV-VIS Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and Agilent 2100 Bioanalyzer using the Bioanalyzer DNA High sensitivity (Agilent Technologies), respectively. The DNA libraries were mixed in groups of 8 (or 4, 6 and 12 in the pilot study) in equimolar ratios to yield a final concentration of 147 ng/μl of each pooled library. The pooled libraries were hybridized with our custom-designed SureSelect Oligo Capture Library SureSelect (Agilent Technologies; SureSelect Human X Chromosome Demo Kit was used for the pilot study (Agilent Technologies)) for 24 h according to the manufacturer's instructions. After incubation, the selected hybrids were purified using magnetic beads and desalted with Qiagen minElute PCR purification column. Post-hybridization amplification PCR with standard primers from SureSelect Target Enrichment System kit and Herculase II Fusion DNA Polymerase (Stratagene, Agilent Technologies) was performed with a denaturation time of 30 s at 98 °C, followed by 18 cycles of denaturation at 98 °C for 10 s, annealing at 57 °C for 30 s and extension at 72 °C for 30 s. The final extension was performed at 72 °C for 7 min. After purification, DNA quantity and quality was checked. A 75-nucleotide (nt) single-end run on the Illumina GAIIx Genome Analyzer (Illumina Inc., San Diego, CA, USA) was performed following the manufacturer's recommendations.

SNP selection and bait design

SNPs were selected to cover all known and putative clinically relevant variations with regard to childhood ALL treatment (Figure 1). First, a list of clinically relevant genes and SNPs was curated, and their influence (known and suspected) in terms of effect on metabolism, transport or drug targets interactions for the 13 most administered chemotherapeutic drugs and their clinical consequences in childhood ALL were evaluated.[2] To extend the list of genes/proteins connected to these drugs, drug–protein associations from different sources such as DrugBank (version 2008)[8] and PharmGKB (version 2008)[9] were gathered. The resulting protein drug targets (from binding data), metabolizing enzymes and drug transporters were integrated into the previous list.

The list of clinically relevant genes and SNPs was further expanded by including their known first-order protein–protein interaction partners using a high-confidence human interactome and other genes participating in the same pathways.[10] This approach allows for investigating complex effects, where several SNPs (potentially in different genes) could exert the same effect through a common mechanism, for example, affecting the same pathway. A list of 969 genes was generated and all genomic variations associated with those genes were extracted using the Ensembl API version 57 (Ensembl, EBI and WTSI, Hinxton, UK) based on the Single Nucleotide Polymorphism Database (dbSNP) 130. SNPs with potential functional impact on the genes of interest were chosen by selecting SNPs resulting in amino-acid changes or frameshifts, SNPs in annotated regulatory regions, variations affecting a stop codon or a splice site, as well as variations within non-coding genes and within mature microRNAs. Furthermore, the list was expanded by including SNPs, which could potentially disrupt predicted microRNA target sites of the listed genes found in Patrocles database.[11]

Baits for the SureSelect Target Enrichment System were designed for all identified SNPs. Each variation was targeted by two baits with a 50% overlap, where the variation was positioned exactly in the middle of the overlap region (Supplementary Figure 1). The physical and cross-hybridization properties of the baits were explored using the oligonucleotide design software OligoWiz[12] and sequence-matching tools



List	Number of SNPs
SNP list	364
Core gene list	7,489
Extended gene list	14,744
Pathway	4,422
Drug list	5,464
miRNA target sites	3,672
Total SNPs	**25,602**
Tiling baits	
Genomic region	CDKN2A, GSTM1, GSTT1
All exons	ETV6, MTAP, OPRM1

Figure 1 SNP selection and bait design. The list of curated genes and drug targets was expanded with the systems biology approach, and known SNPs in these genes were selected with potential functional significance. The baits targeting selected SNPs were designed to minimize the extent of cross-hybridization and self-folding of the baits, as well as extreme levels of GC content.

BLAST[13] and SeqMap.[14] The extent of baits prone to cross-hybridization, self-folding, extreme levels of GC content or baits targeting highly variable regions, which could decrease specificity or efficiency of the baits, was limited to a minimum. However, some potentially problematic baits were still included in the design because of the clinical importance of their target region (Supplementary Figure 2). To exploit the whole capacity of the method and to probe for key deletions, additional baits were designed tiling all the exons of *ETV6, MTAP* and *OPRM1,* and the entire genomic regions of *CDKN2A, GSTM1* and *GSTT1*, allowing full sequencing of these genes of particular importance in childhood ALL. The baits tiling the genomic regions of the drug-metabolizing genes *GSTM1* and *GSTT1* were used to detect the deletion state of those genes. To estimate copy number, a depth ratio was calculated from the number of reads in the targeted genomic region normalized by size of the region and total number of reads for the sample. The final design included baits targeting 25 602 clinically relevant SNPs, as well as 1200 baits targeting the exons or the genomic regions of the above-mentioned genes, with specific impact on the treatment outcome in childhood ALL (Figure 1).

Barcodes

In the pilot study, 12 different barcodes (Supplementary Table 2), with the last base being a thymidine (T) necessary for ligation to chromosomal DNA fragments with a 3′ adenosine (A) overhang, were tested. These four base barcodes in combination with 75 nt sequencing reads render high quality, unambiguously mapped reads, while using only 5% of the read length for sample identification.

Data analysis

The high-quality reads obtained from sequencing were aligned to the NCBI37 reference human genome (version GRCh37) using the Burrows–Wheeler Alignment Tool.[15] The alignment was refined by means of quality score recalibration and around indel realignment using Genome Analysis ToolKit package.[16] SNP calling was performed with SAMtools package[17] using default settings. The threshold set for SNP calling was minimum 10x sequencing depth; however, 4x was also accepted as a threshold for high-priority SNPs when 10x depth was not available. The data was further analyzed with help of SAMtools and BEDtools[18] packages and custom-written Perl scripts.

Results

Pilot study

The applied baits from the SureSelect Human X Chromosome Demo Kit were designed to capture 85% of the human X chromosome exons, tiling exons with a 50% overlap between consecutive baits. In an attempt to reflect the application of this method to custom-genotyping purposes, the performance was assessed on a set of target positions defined as base pairs positioned exactly in the middle of the 50% overlap region of two consecutive baits (Supplementary Figure 1). The average sequence depth achieved with this approach is higher and more uniform in the immediate surrounding of the target position when compared with placing the target position in the middle of a single bait.

The results of the pilot study showed that average sequence depth decreases with an increasing number of pooled samples; however, the distribution of reads is well balanced and sufficient for genotyping even with 12 pooled samples (Supplementary Figure 3). Each sample had between 55 and 65% sequencing reads mapped to the targeted regions and between 65 and 75% of the reads mapped to targeted regions plus/minus 100 bp of target (Supplementary Figure 4), indicating that sample competition during hybridization was not an issue. An average depth of at least 10x was achieved for ~88% of the target positions, with a standard deviation of 10% when pooling 12 samples (Supplementary Figure 5). Increased pooling affected the number of detected target positions. However, depending on the desired depth of sequence reads and required target size (number of SNPs for genotyping studies), the amount of pooling can be adjusted appropriately. Genotyping using depths from 4x to 12x has recently been tried[19,20] and consensus for 10x SNP calling is relatively high, while mapping against a known reference genome such as the human.

SNP calls from the targeted sequencing approach were compared with genotype calls using in-house Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA) and public HapMap data. The comparison was restricted to all probes on the array matching SNP calls within targeted regions from the sequencing kit. The accuracy of the multiplexed targeted sequencing approach was assessed on 12-pooled samples based on concordance for sequencing and array-based genotype calls. SNP calls with the coverage depths of 4x, 10x and 20x were compared with the SNP calls obtained from the array, and the percent concordance for each sample was calculated as the percent of calls that were the same for each depth (Supplementary Figure 6). As expected, the number of compared SNPs decreased with increasing depth required; however, it did not affect the concordance rates significantly. In this study, calling SNPs at 4x depth appeared acceptable when 10x depth was not available. All samples achieved concordance well above 95% at 10x coverage depth (the same holds for 4x coverage, with an exception of the CGCT labeled sample). Examining regions that had 20x or better sequence depth, 9 out of the 12 samples were 100% concordant; however, the number of SNPs decreased by ~38% because of minimum sequencing depth requirement (20x) as compared with 10x coverage depth.

Childhood ALL samples

Sequencing of the 48 samples generated a total of 39 Gb of data in FASTQ format and 32 Gb passed the default Illumina quality filter. Out of those, 23.2 Gb were uniquely mapped to the reference genome (Table 1). On average, 53% (sd of 9%) of the high-quality sequencing reads were mapped to the target regions, which shows that the distribution of reads was relatively balanced. In addition, on average, 94% of the targeted SNPs were covered at least with one sequencing read for each sample, and ~73% of those SNPs achieved at least 10x sequencing depth. The average sequence coverage for the covered variations from the list of targeted SNPs was 23x across all samples. The average depth coverage for the targeted exons of *ETV6, MTAP* and *OPRM1* genes was 32x, whereas the average coverage for the genomic region of *CDKN2A* gene was 31x. Average depth coverage for *GST*s could not be estimated, as it varies with the deletion state of the genes.

Bait design

The performance of capturing baits depends on their physical and cross-hybridizational properties (Supplementary Figure 2). Baits used in the design were explored using the OligoWiz program for their probability of self-folding and cross-hybridiza-

tion for the presence of low-complexity regions and their GC content. All the tested parameters seem to influence bait performance; therefore, these properties should be taken into consideration during bait design. If required, higher depths for regions of particular interest can be obtained by targeting them with higher numbers of overlapping baits. In this study, for a list of high-priority SNPs with known significant impact on the treatment outcome in childhood ALL,[2] four different baits (instead of two) have been designed for each SNP. The obtained average coverage for those was 35x, as compared with the average coverage of 23x for the whole list of SNPs. Special care must be taken when targeting regions on a human mitochondrial genome, as these DNA fragments are overrepresented in a genomic DNA sample. Mitochondrial regions will attract significantly more reads, and might therefore dominate the sequencing results. In our study, 66 baits targeting regions on the mitochondrial genome were included, achieving an average depth of 4350x and the reads corresponding to those constituted 15% of all the mapped sequencing reads.

Genotype validation

Genotype data from the 48 childhood ALL samples on seven SNPs and two gene deletions were used for validation of the sequencing results (Table 2). Patients were previously genotyped for *CYP3A5*3* 6986A>G (rs776746), *RFC1* 80G>A (rs1051266), *TPMT*3B* 460G>A (rs1800460), *TPMT*3C* 719A>G (rs1142345), *MTHFR* 677C>T (rs1801133) and *MTHFR* 1298A>C (rs1801131) by allelic discrimination[5,6] (and unpublished data). *GSTP1* 313A>G (rs1695), *GSTM1* and *GSTT1* deletions were genotyped by multiplexing PCR, which simultaneously detects *GSTT1* and *GSTM1* gene copy number and *GSTP1* 313A>G.[21] Based on the calculated coverage depth ratio for the genomic regions of *GSTM1* and *GSTT1,* it was possible to distinguish three distinct clusters corresponding to homozygous deletions, heterozygous deletions and the wild type (Figure 2). Concordance for the seven SNPs was between 91 and 100%, and 100% for the two gene deletions (Table 2).

Table 1 Summary of the sequencing of the childhood ALL samples

Total samples	48
Total sequencing lanes	6
Total samples per lane	8
Total raw reads (Gb)[a]	39
Total high-quality reads (Gb)[a]	32
Total mapped reads (Gb)[a]	23.2
Mean percent on target (%)	53
Mean mapped depth on target (x)	35
SNPs with read depth (%)	
⩾1x	94
⩾4x[b]	90
⩾10x[b]	73
⩾20x[b]	48
Mean variant SNP sites per individual (at ⩾10x)	3896
Novel (not in dbSNP130; %)	9.5
Mean variant indel sites per individual	1024

Abbreviations: ALL, acute lymphoblastic leukemia; SNP, single-nucleotide polymorphism.
In all, 94% of the targeted SNPs were covered by at least one read; of these, 73–90% could be called at 4x or 10x depth. With eight samples per capture and per sequencing lane, the per sample cost was in the $200–300 range at 2010 prices.
[a]Files in FASTQ format.
[b]Of SNPs covered by at least one read.

Discussion

In recent years, clinically important genetic variations have been thoroughly investigated in childhood ALL.[2] Even though minimal residual disease monitoring[22] and extensive toxicity scoring have been established in some treatment protocols,[1] still very few groups include genetic variations in their treatment strategies.[6] This primarily reflects the lack of extensive targeted analyses of genetic variants and the costs associated with such analyses. Multiplexing before capture, target enrichment and sequencing allows screening of ~25 000 custom-selected SNPs simultaneously and could therefore be a solution. As shown in the pilot study, pooling of 4–12 samples in a single lane of an Illumina GAIIx with 75 nt single-ended reads is sufficient to generate 10–20x sequence depth over more than 80% of the target region per sample (Supplementary Figure 3). If a 10–20% drop in coverage (or target size) is allowed, pooling 6 or 12 samples can easily be achieved. Pooling of up to 12 samples showed relatively balanced results, but larger dispersion at higher depth. To reduce this variation, we hypothesized that in case of pooling 8–10 samples, an average sequence depth of 10x for ~80% of the intended regions could be expected. The amount of pooling was therefore adjusted to eight samples at a time because of the large size of the desired target and need for high sequence depth. The next-generation sequencing technology is rapidly evolving; hence, in future studies, it will be possible to either pool more samples or obtain higher coverage. It was observed in different sequencing runs that not all barcodes performed equally well. For example, we found that barcode CGCT had a lower performance compared with the rest (Supplementary Table 1, Supplementary Figures 5 and 6). This could be inherent to the barcode itself, but also influenced by experimental variation (for example, small differences in the amount and quality of the different adapters and variation in library preparation). Additional studies exploring this and the individual barcodes could provide information to further improve the study design.[23]

Table 2 Genotype concordance for the 48 childhood ALL samples analyzed by single PCR and by multiplexing sequencing

	*CYP3A5*3 6986A>G*	*TPMT*3B 460G>A*	*TPMT*3C 719A>G*	*RFC1 80G>A*	*MTHFR 677C>T*	*MTHFR 1298A>C*	*GSTP1 313A>G*	*GSTM1 deletion*	*GSTT1 deletion*
Single PCR/sequencing (*N*)	48/47	30/48	30/47	46/48	29/46	29/48	42/48	42/48	42/48
Analyzed by both methods (*N*)	47	30	29	46	28	29	42	42	42
Concordant (*N*)	46	28	28	42	26	28	42	42	42
Concordance (%)	98	93	97	91	93	96.55	100	100	100

Abbreviation: ALL, acute lymphoblastic leukemia.

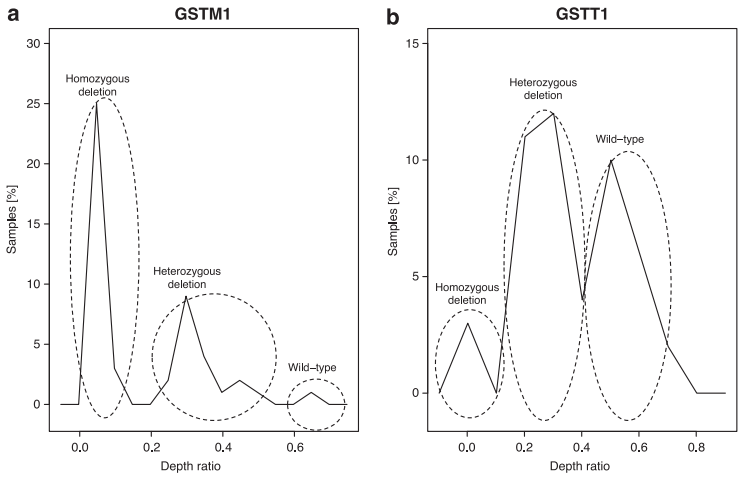


Figure 2 Copy number detection in sequence data. The state of gene deletion is assessed based on the distribution into three distinct clusters corresponding to the wild-type, heterozygous and homozygous deletions. Depth ratio was calculated from the number of reads in the targeted genomic region normalized by size of the region and total number of reads for the sample. Results were validated for 42 of the 48 samples using multiplexing PCR and showed 100% concordance. (**a**) Deletion state for the *GSTM1* gene. (**b**) Deletion state for the *GSTT1* gene.

We here demonstrate that it is possible to reliably genotype childhood ALL patients for a large number of SNPs simultaneously using multiplexed target enrichment, followed by sequencing. Sequencing data for ~94% of the targeted SNPs for each sample was obtained, and for 73% of those, the achieved coverage depth was sufficient for high-confidence genotype calling (at least 10x). In addition, this methodology is easy to adapt in a sequencing lab and has a low entry level (80–100 samples), thus allowing redesign of content during the course of large sample projects. In a single design, exon tiling can be combined with SNP or somatic-mutation detection, and we show here that copy number can be reliably inferred through sequencing. Whole-exome sequencing is gaining popularity among targeted sequencing efforts as a way to improve sequence depth and reduce cost compared with whole-genome sequencing, as exons span only ~1.2% of the whole human genome. This is a valuable approach and it has recently been shown that some multiplexing (3–5 samples) can be accomplished.[24] However, when sequencing the exons only, many regions are missed that may have important biological functions such as transcriptional or translational regulation of the protein-coding sequences. Many studies indicate non-coding SNPs to be clinically relevant; therefore, it is crucial for pharmacogenetic studies to also investigate regions outside of the protein-coding parts of the genome. We demonstrate that a more targeted hypothesis-driven panel can be constructed, assayed reliably and at much lower costs than exome sequencing. Pooling of eight patient samples before capture reduces the costs of the capture library, hybridization reagents and, none the least, the costs of sequencing by eight times. Furthermore, the patients are genotyped for 25 000 targeted SNPs simultaneously, reducing the cost per SNP per patient even further compared with conventional methods.

The presented method for performing high-throughput, low-cost, customized genotyping will allow wider application of studying clinical impact of genomic variations. Immediate applications include validation of GWAS, assaying somatic mutations or a panel of SNPs such as in drug toxicology studies. Furthermore, the flexibility of the bait design enables researchers to adapt the SNP content as new knowledge emerges. Future applications of this method in upcoming childhood ALL studies will move us closer to pharmacogenetic-based personalization of therapy in childhood ALL—and possibly other cancers. Such studies are currently ongoing in the NOPHO Study Group.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to the patients who participated in the study and their referring physicians. We thank Kirsten Kørup Rasmussen for very helpful technical assistance and Jannie Gregers for providing us with previously generated SNP data. We acknowledge The Technical University of Denmark Multi-Assay Core for providing technology consultation and laboratory resources. AW, MDD and LB analyzed, interpreted data and wrote the manuscript. AW performed the sequence analysis. MDD, LB, HL, KS and RG designed the experimental research project setup. LB, MDD, LRH and BFN performed the experimental work and the Affymetrix 6.0 SNP Arrays. RG performed data analysis supervision. MB and NT performed the Illumina sequencing. KA, LG, TSP and NW performed parts of the data analysis. JN provided cell lines. LG, HL, RG, SB and KS provided critical input to the project and manuscript. This study was supported by grants from The Danish Cancer Society (Grant numbers R2-A56-09-S2 and R20-A1156-10-S2), The Danish Childhood Cancer Foundation, The Otto Christensen Foundation, The Villum Kann Rasmussen Foundation, The Ministry of Health (Grant number 2006-12103-250), The Novo Nordisk Foundation, The Danish Research Council for Health and Disease (Grant numbers 271-06-0278, 271-08-0684), The University Hospital Rigshospitalet, Denmark, The Lundbeck Foundation, The research program of the UNIK: Food, Fitness and Pharma for Health and Disease, The Danish Ministry of Science, Technology and Innovation and The Wilhelm Johannsen Centre

for Functional Genome Research that is established by the Danish National Research Foundation.

References

1 Schmiegelow K, Forestier E, Hellebostad M, Heyman M, Kristinsson J, Soderhall S *et al*. Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia* 2010; **24**: 345–354.
2 Davidsen ML, Dalhoff K, Schmiegelow K. Pharmacogenetics influence treatment efficacy in childhood acute lymphoblastic leukemia. *J Pediatr Hematol Oncol* 2008; **30**: 831–849.
3 Schmiegelow K, Al-Modhwahi I, Andersen MK, Behrendtz M, Forestier E, Hasle H *et al*. Methotrexate/6-mercaptopurine maintenance therapy influences the risk of a second malignant neoplasm after childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study. *Blood* 2009; **113**: 6077–6084.
4 Lund B, Åsberg A, Heyman M, Kanerva J, Harila-Saari A, Hasle H *et al*. Risk factors for treatment related mortality in childhood acute lymphoblastic leukaemia. *Pediatr Blood Cancer* 2011; **56**: 551–559.
5 Gregers J, Christensen IJ, Dalhoff K, Lausen B, Schroeder H, Rosthoej S *et al*. The association of reduced folate carrier 80G>A polymorphism to outcome in childhood acute lymphoblastic leukemia interacts with chromosome 21 copy number. *Blood* 2010; **115**: 4671–4677.
6 Schmiegelow K, Forestier E, Kristinsson J, Soderhall S, Vettenranta K, Weinshilboum R *et al*. Thiopurine methyltransferase activity is related to the risk of relapse of childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study. *Leukemia* 2009; **23**: 557–564.
7 Relling MV, Yang W, Das S, Cook EH, Rosner GL, Neel M *et al*. Pharmacogenetic risk factors for osteonecrosis of the hip among children with leukemia. *J Clin Oncol* 2004; **22**: 3930–3936.
8 Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006; **34** (Database issue): D668–D672.
9 Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB *et al*. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002; **30**: 163–165.
10 Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O *et al*. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* 2007; **25**: 309–316.
11 Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res* 2010; **38** (Database issue): D640–D651.
12 Wernersson R, Juncker AS, Nielsen HB. Probe selection for DNA microarrays using OligoWiz. *Nat Protoc* 2007; **2**: 2677–2691.
13 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–410.
14 Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* 2008; **24**: 2395–2396.
15 Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010; **26**: 589–595.
16 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
17 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
18 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.
19 Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H *et al*. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010; **42**: 969–972.
20 Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
21 Buchard A, Sanchez JJ, Dalhoff K, Morling N. Multiplex PCR detection of *GSTM1, GSTT1*, and *GSTP1* gene variants: simultaneously detecting *GSTM1* and *GSTT1* gene copy number and the allelic status of the GSTP1 Ile105Val genetic variant. *J Mol Diagn* 2007; **9**: 612–617.
22 Campana D. Progress of minimal residual disease studies in childhood acute leukemia. *Curr Hematol Malig Rep* 2010; **5**: 169–176.
23 Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ *et al*. Identification of genetic variants using barcoded multiplexed sequencing. *Nat Methods* 2008; **5**: 887–893.
24 Nijman IJ, Mokry M, van BR, Toonen P, de BE, Cuppen E. Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nat Methods* 2010; **7**: 913–915.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)

Chapter 7

Integrative variation analysis

This chapter presents some of the integrative analyses methods applied in Papers III, IV and V, as well as the unpublished analysis of the individual disease risk based on genome sequence of an Aboriginal Australian.

7.1 Single SNP associations

Conventionally in association studies involving cases and controls one compares the allele frequencies at all the genotyped loci between the two groups of individuals to determine whether there is a statistically significant association between the genotype at a given locus and the phenotype under investigation. There exist several models of genetic penetrance of the phenotype, including additive, multiplicative, recessive and dominant, which determine the risk of developing the phenotype dependent on the present number of risk alleles. Testing for association with phenotype is usually performed for each SNP separately by means of chi-squared test, which tests independence of the rows and columns of a contingency table of counts of phenotype status by genotype or allele count. As chi-squared test is an approximation of the results of a Fisher's exact test, the latter can also be used for association studies and is the recommended choice for small sample sizes or in cases when cell count less than 5 can be expected. When it is necessary to adjust the association for possible confounding covariates like environmental factors or population stratification a logistic regression model can be used, and if the phenotype is a continuous trait a linear regression model can be applied. After performing multiple single SNP association test, one should correct the significance values for multiple testing to avoid incidental findings.

7.2 Rare variant accumulation

Reasoning behind conducting GWA studies included the "common disease common variant" hypothesis suggesting that common diseases are caused by multiple common variants. However, when the majority of GWAS findings turned out to be associations with only modest effect sizes and explaining only small fraction of heritability, the focus has been shifted to investigation of rare variants [78]. This gave rise to the "multiple rare variant" hypothesis suggesting that common inherited diseases are caused by aggregation of multiple rare variants with larger effect sizes [13]. Different scenarios of common and rare variant attributing to phenotype are illustrated in Figure 7.1. The studies of rare variants have recently been made possible due to a continuing decrease in NGS technologies cost and completion of 1,000 Genomes Project serving as a public rare variant catalogue.

There exist several methods for investigating effects of rare variants, which can be summarized as: single marker tests, multiple marker tests, collapsing methods and approaches based on similarities among individual sequences reviewed in an article by Bansal *et al.* [10]. While single marker tests like Fisher's exact test can be applied to study rare variants, they are usually underpowered even in very large samples and alternative approaches considering multiple markers are required. Even though the individual frequency of rare variant is low, collectively their abundance makes them quite common, occurring on average every 17 bases as reported by a recent study by Nelson *et al.* [86]. This is the assumption for collapsing methods, which simply test for accumulation of rare minor alleles in the case group at a specific associated genomic site. This method was applied to study effects of multiple variants in Paper III by performing the cumulative minor-allele test (CMAT) [136] on the coding SNPs grouped by their genic regions or protein-protein complexes. The obvious limitation of this type of methods is that investigating accumulation of rare variants requires large sample sizes and the result is largely influenced by the size of investigated genomic region. Alternative methods include inspecting the local similarity of the DNA sequences between individuals motivated by phylogenetic analyses or multiple regression methods allowing for adjusting for additional predictor variables like covariates or common variants. Overall, numerous methods for studying the effects of rare variants exist, however there is no consensus on which of them are most applicable and what is their power and robustness in different contexts. Despite that, increasing amount of studies report and discuss the importance of rare variants, including a recent study demonstrating that rare variants in *SLCO1B1* have greater effects on methotrexate clearance in childhood ALL patients than common variants and thereby showing that rare variants affect traits beyond the disease susceptibility [101].

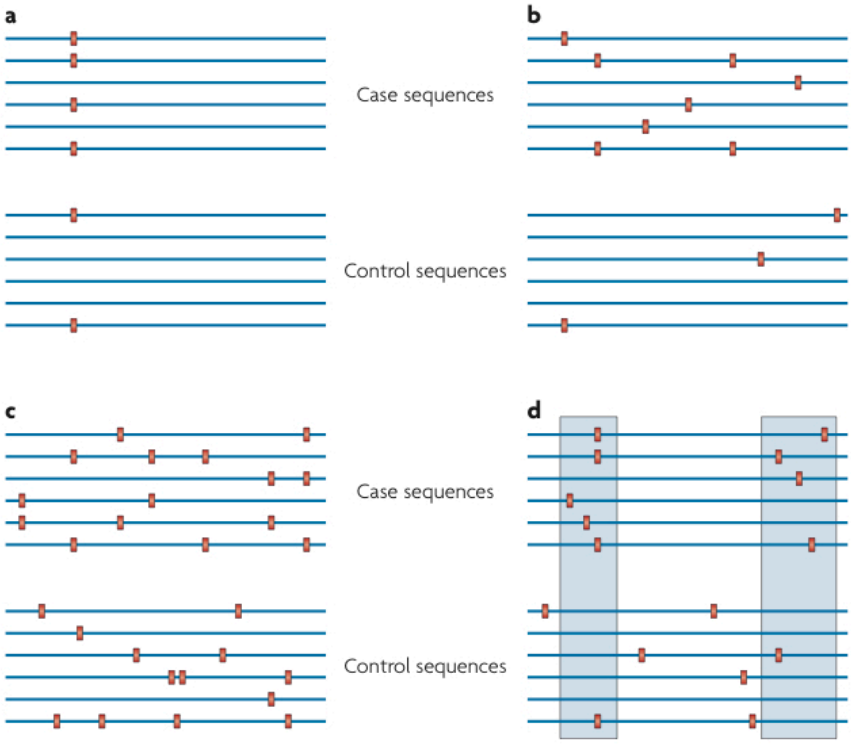


Figure 7.1. Scenarios in which DNA sequence variants distinguish cases and controls. The blue lines indicate genomic regions; red boxes indicate variants. a | Variants at a single locus with common alleles are more frequent in cases then controls. b | Multiple rare variations contribute to the phenotype such that the collective frequency of these variations is greater in cases. This would create a greater diversity of haplotypes or DNA sequences among the cases. c | Multiple rare variations contribute to the phenotype but act in a synergistic fashion, such that cases are likely to have more similar DNA sequences compared to controls. d | Multiple rare variations contribute to a phenotype but the variations contributing to the phenotype reside in specific genomic regions. This situation would create greater sequence diversity among the cases, as in part b, but only in the relevant genomic regions. Source of figure and legend: Bansal *et al.* 2010 [10].

7.3 Pathways analysis

While many association studies concentrate on detecting single marker association signals, it is important to remember that genes do not act alone. The interactions between various cellular components such as genes, proteins, small RNAs or metabolites create complex networks, which together affect the phenotypes. Often perturbation of different interacting components of the same network result in disruption of the same molecular mechanism and give rise to the same observable trait. While studying associations of genomic variations to a phenotype, one can examine the collective effects of multiple SNPs acting in the same biological pathway, serving as a simplified model of the complex cellular interactions. This kind of integrative analysis may yield more robust results and is more sensitive to detect a sum of modest but not independent effects in a group of individuals. Finally, pathway analyses can aid translation of the findings into clinical setting, as other pathway members might be more suitable drug targets than the most associated gene.

Several methods have been developed to study pathway-mediated effects in GWA studies inspired by the success of gene-set enrichment analysis for gene expression data [122]. The most popular of those methods are described in a review by Wang *et al.* 2010 [130]. One of the major difficulties in conducting pathway-level associations for GWAS is mapping the SNPs to their respective genes, as most of the variations included on the widely-used SNP arrays reside in non-coding regions. Further, the amount of genotyped SNPs is too large to include in the analysis, so different search space reducing techniques are used, including filtering out SNPs in high LD or only considering one SNP per gene with the lowest p-value. These practices may lead to loss of information, as in the case when association tests only consider the individual SNPs p-values rather than the SNP genotype. On the other hand, limiting the investigations to only the most significant SNPs can lead to missing effects mediated by multiple genetic variants acting in a concert but individually having only small effects. So far pathway analyses for NGS based studies are rather limited and concentrate mostly on detecting overrepresentation of mutations within defined pathways without taking into account the more common variants.

This was the motivation behind the pathways association analyses presented in Papers III and IV in this thesis. Due to unique design of the assay used in these studies, the number of genotyped SNPs was much lower than in a typical GWA study and the assayed SNPs were selected to reside in coding or regulatory regions, hence the biggest difficulties of GWAS based pathway analyses were not an issue in this case. In these two papers the investigated SNPs have been grouped by all pathways from Reactome database [31] and eleven drug metabolic pathways from PharmGKB [54] relevant to drugs administered in childhood ALL. Only SNPs observed at a minimum MAF of 0.5 % and likely to affect the protein function were included: non-synonymous coding, frame-shift coding, as well as SNPs affecting the stop

codons and splice sites. The association testing of groups of SNPs was performed by artificial neural network training with three-fold cross-validation to avoid overfitting. All the SNPs were encoded by 3 values corresponding to likelihood of each genotype derived from the VCF file produced during the SNP calling step. In case of lack of sequencing coverage at a given position the values used to encode the SNP were the observed frequencies of the genotypes in remaining samples. Since it is not computationally feasible to train neural networks on all possible combinations of SNPs, even when divided into smaller sets defined by pathways, the selection of informative SNPs for each pathway followed a simple two step procedure: 1) all the possible combinations of up to 3 SNPs were tested for each pathway and 2) the combinations were iteratively increased up to 15 SNPs by adding another SNP to top 20 previous combinations of SNPs. The performance of neural networks models was assessed by Matthew's correlation coefficient (MCC), which takes into account the numbers of true and false positives and negatives. MCC is calculated according to the following formula:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where:
TP = number of true positives
TN = number of true negatives
FP = number of false positives
FN = number of false negatives.

This approach validated the hypothesis presented in Paper III that inter-individual variation in drug disposition is largely effected by the host genomic background at pharmacogenomic domains. For both investigated phenotypes (MRD levels after remission induction therapy and risk of relapse) pathways comprising of transporter genes were found among the top associated collections. Additionally, using this method the most important biological mechanisms contributing to phenotypes can be identified and used together with other predictive variables in prediction of the treatment outcome.

Pathway-level analysis is becoming a popular method for complementing single SNP associations, however there are several challenges to be faced. Studying phenotypes by means of pathway associations has the obvious drawback that many human genes are not studied well enough to be included in known pathways, so the effects of variations of those genes would not be accounted for in pathway analysis. Moreover, the pathway representations collected by public databases are often not complete and are limited by current understanding of the biological processes. Lastly, different pathway resources contain different descriptions of the same biological pathways differing in levels of complexity and numbers of included components. Therefore, comprehensive, consistent and reliable pathway resources are needed in

the future to conduct unbiased association analyses on pathway level.

7.4 Individual disease risk

GWA studies provide an excellent way of identifying variants contributing to disease susceptibility. This information can be further used to predict the susceptibility to a given disease for an individual and is used for instance by direct-to-consumer genetic testing provided by companies like 23andMe, deCODEme or Navigenics as described in Chapter 1.4. The usefulness of combining genomic information together with patient's clinical characteristics for disease risk estimation and genetic counselling was also demonstrated by Ashley *et al.* 2010 [7].
Motivated by that, we have attempted to estimate the disease risk susceptibility of an aboriginal Australian sequenced in a study by Rasmussen *et al.* 2011 [102]. The genome was sequenced from a lock of hair collected a century ago, therefore no additional information of the health status or family disease history was available for the individual and the disease risk prediction had to be based solely on genomic information. The GWA studies are usually performed on individuals with European ancestry, therefore their findings might not be directly applicable to studying disease risk of a member of a much different isolated population of aboriginal Australians. However, the overall disease mechanisms and associated genes are not expected to differ between the populations, therefore this analysis concentrated on known genes associated to diseases known to be more prevalent in the aboriginal Australians.
The burden of disease suffered by Australians with aborigine ancestry is estimated to be two-and-a-half times greater than that of the entire Australian population [11]. The increased risk is seen for a variety of disorders including cardiovascular, respiratory and kidney diseases, as well as diabetes and different types of cancer. To study the disease risk of the aborigine we examined the genetic variant enrichment in a set of functional modules associated with diseases. The investigated collections of genes were compiled using KEGG [92] and Reactome [31] pathways, Gene Ontology categories [6], OMIM [50], GeneCards [104] and protein-protein complexes, including disease associated protein-protein complexes curated from phenome-interaction networks [68, 67]. Functional modules not appearing significant in other modern day genomes were prioritized. The enrichment score was calculated based only on the high-confidence missense and nonsense variations, since those have the most potential for functional effects. The simple assumption here is that genetic changes causing functional, especially deleterious changes, to a protein in a disease associated complex or network, would impact disease development. Enrichment of such functional effects in a complex increases the likelihood of impact. The scoring used here was very similar to gene set enrichment techniques used in prior literature [36, 38, 95, 122]. Each disease-associated set of genes was scored using hypergeometric testing following the formula:

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

where:
k - number of genes related to a given disease in which aborigine had potentially deleterious variants
m – total number of genes in a given disease module
n – total number of genes in which the aborigine had potentially deleterious variants
N – total number of genes included in all the considered disease modules.
All numbers were corrected for sequencing coverage, e.g. genes with no sequence coverage were not included in numerator or denominator.

The final enrichment score was calculated as *S=1-log (*P*)*, where P was derived from the hypergeometric distribution. These enrichment scores were plotted as the length of the disease cones in the half-circular plot (Figure 7.2).

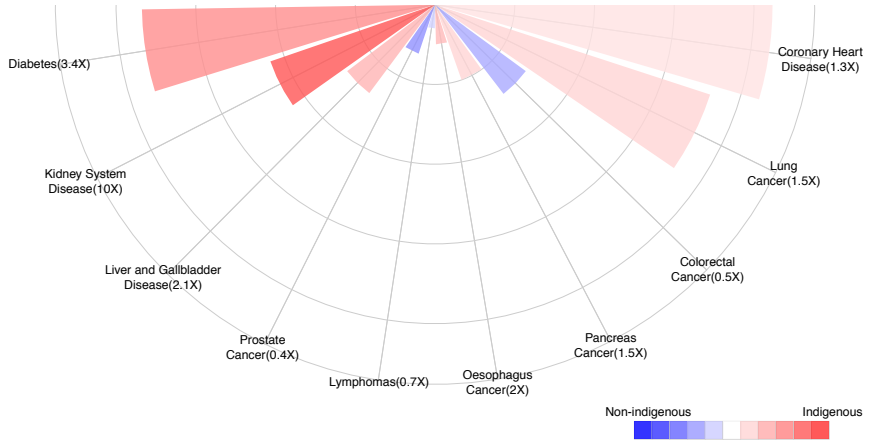


Figure 7.2. Disease risk for aborigine.

To bring the P-values to a graphable scale, -log (P-value) was used and the output was linearly rescaled to lie between 0 and 1. To prioritize diseases to display, prior knowledge from the Australian Bureau of Statistics and the Australian Institute of Health and Welfare [11] was used. The ratio of age-standardized incidence in people of Aboriginal Australian descent relative to those of non-Aboriginal Australian background was used as a measure of ”interestingness” to select diseases to focus on. Overall, diseases with a higher occurrence in indigenous Australians, like diabetes, lung cancer and

coronary heart disease, also obtained a higher enrichment score in the associated disease complexes. Accordingly cancer types with a less increased risk in indigenous Australians, like lymphomas, colorectal cancer or prostate cancer, show lower enrichment scores.
A single aborigine genome does not have the statistical rigor of genome wide disease association studies. However, it can still pinpoint areas of further research interest, especially as the genome shows no signatures of recent admixture. Disease risk inference has previously been shown to be successful even when applied to a genome of a single individual [7]. This kind of analysis could be performed having the results of various association studies performed in the Aboriginal Australian population, or one closely related to it, otherwise the results cannot be directly compared. The findings of this study have also demonstrated that rare and novel mutations are abundant in the genes known to be associated with disease. Therefore even without knowing the exact causative variants in a given population, we can use missense variants accumulation analysis to identify subjects at risk of showing severe phenotypes, using an approach similar to strategies applied before in other studies [62]. This kind of analysis goes few steps beyond the current state of art in this field. A recent paper by Fujimoto *et al.* [43] shows enrichment of the functional polymorphisms in olfactory receptors by gene ontology categories enrichment analysis. However these findings lack comparison to other modern genomes, which were found in this analysis to show similar enrichment patterns. This stresses the necessity to compare the trends across the genomes, therefore in this work the modules not appearing to be significantly enriched in other modern genomes were prioritized, and prior knowledge of disease incidence ratio was included to verify the findings. This approach also integrates different sources of gene-phenotype information making the analysis more comprehensive. Similar approaches have been applied before to rank overrepresented sets of genes [36, 38, 95, 122], however this kind of analysis has not been applied before to genome wide genotype data in context of inferring a disease risk.

7.5 Subgrouping patients

Copy number variations are another common type of genomic variation and can contribute to disease susceptibility. Changes in copy number have been implicated in susceptibility to among others inflammatory autoimmune disorders [2], autism and schizophrenia [27], however the most common CNVs have been shown to not play a significant role in causing disease [30]. On the other hand copy number alterations (CNAs) are commonly acquired in cancer and by altering the expression of involved genes might contribute to cancer heterogeneity between patients. In Paper V in this thesis we investigated CNAs in a subgroup of childhood ALL cancers with a *t(12;21)* translocation. This group of patients is characterized by very good prognosis, however cases of relapse or other events can also occur in patients with this translocation. This observation leads to hypothesis that *t(12;21)* childhood ALLs

should not be treated as a homogeneous group and different subgroups can be distinguished within those patients with different prognosis.

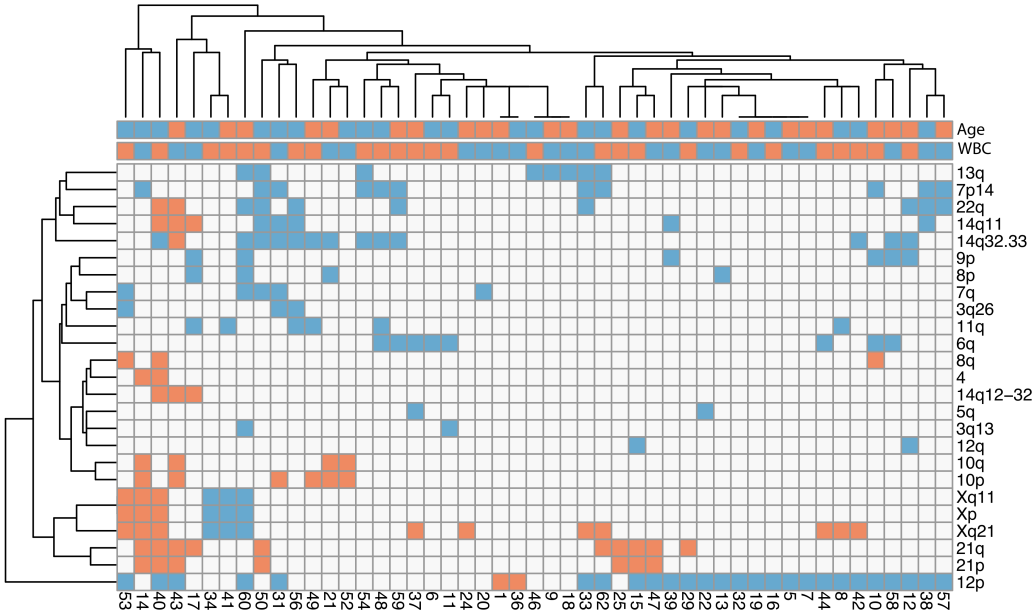
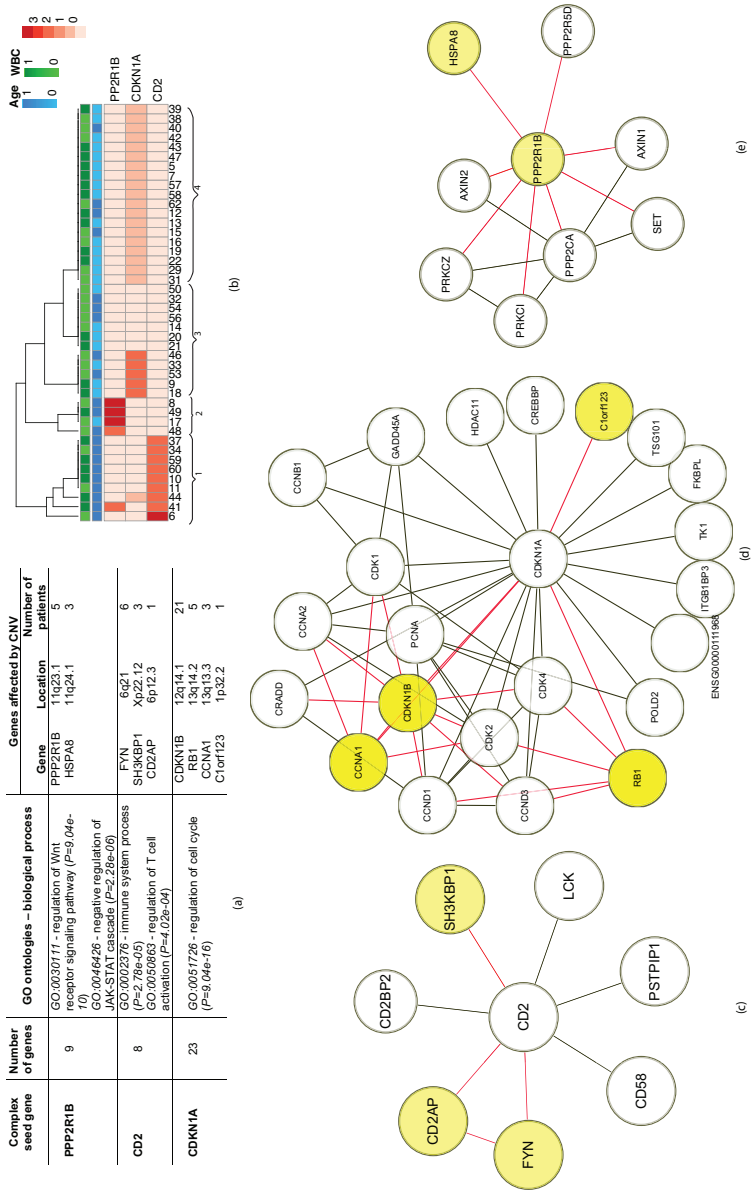


Figure 7.3. Heatmap clustering of patients and their recurrent aberrations summarized in cytoband, chromosome arm or whole chromosome regions. Source: Supporting Figure 1 from Paper V.

Conventional heatmap analysis of recurrent aberrations did not reveal any obvious patterns of patient subgrouping (Figure 7.3), therefore the detected CNAs were mapped to protein-protein interaction complexes which they are likely to affect to identify specific biological mechanisms affected. The performed analysis was limited to losses of genetic material as these are likely to result in a direct loss of function and hence are most disruptive to the complexes. All possible combinations of 2-5 different complexes were created and tested to find a set of complexes affected in the largest possible group of patients with the smallest possible overlap between different complexes. The resulting list of top combinations was filtered based on expression data to include only genes highly expressed in B lymphoblasts and further prioritized based on annotations of the participating genes related to leukaemia. The three protein-protein complexes that best discriminated between the patients were seeded from genes *CD2*, *CDKN1A* and *PPP2R1B* (Figure 7.4) and aberrations in those complexes enabled identification of four distinct groups



Complex seed gene	Number of genes	GO ontologies – biological process	Genes affected by CNV: Gene	Location	Number of patients
PPP2R1B	9	*GO:0030111* - regulation of Wnt receptor signaling pathway (*P=9.04e-10*)	PPP2R1B	11q23.1	5
		GO:0046426 - negative regulation of JAK-STAT cascade (*P=2.28e-06*)	HSPA8	11q24.1	3
CD2	8	*GO:0002376* - immune system process (*P=2.78e-05*)	FYN	6q21	6
		GO:0050863 - regulation of T cell activation (*P=4.02e-04*)	SH3KBP1	Xp22.12	3
			CD2AP	6p12.3	1
CDKN1A	23	*GO:0051726* - regulation of cell cycle (*P=9.04e-16*)	CDKN1B	12q14.1	21
			RB1	13q14.2	5
			CCNA1	13q13.3	3
			C1orf123	1p32.2	1

Figure 7.4. Protein-protein complexes distinguishing subgroups of patients. A. Table listing genes affected by CNAs in each complex together with Gene Ontology annotations of the biological process in which the complex is involved. B. Heatmap showing distinct groups of patients identified through protein-protein complex analysis. Sidebars show age and white blood count (WBC) of patients with lighter colours representing values below mean and darker colours represent values above mean. C. Graphical representation of the protein-protein interactions within complexes. Yellow nodes represent genes affected by CNAs in this study. Source: Supporting Figure 2 from Paper V.

of patients with potentially different biological mechanisms underlying the leukaemogenesis.

The three complexes are involved in different biological processes and mechanisms, potentially reflecting the different aetiology of leukaemia in different groups. Briefly, the protein-protein complex seeded from *CD2* is involved in immune system processes, including regulation of T cell activation, and has been associated with various types of leukaemia and lymphoma. The *CDKN1A* complex is mainly involved in regulation of cell cycle and finally the *PPP2R1B* complex plays a role in regulation of the Wnt receptor signalling pathway and in negative regulation of the *JAK-STAT* cascade, both pathways are known to be of high relevance to leukemia. Interestingly, this patient classification also shows some agreement with clinical characteristics of the patients, e.g. patients with deletions in *CD2* complex are generally older and have high WBC, patients with single deletion of *CDKN1B* are young and have high WBC, while patients with no deletions or deletions other than *CDKN1B* in the *CDKN1A* complex and patients with deletions in *PPP2R1B* have low WBC. This observation further strengthens the presented hypothesis and provides additional evidence of patient subgroups observable beyond the CNAs occurrence.

Part III

SNP profiling of treatment efficacy in childhood ALL

Chapter 8

Paper III - Extensive targeted SNP profiling predicts early treatment response and risk of relapse in 864 childhood ALL patients

The overall current cure rate for childhood acute lymphoblastic leukaemia is approximately 80%, thus around 20% of treated children die from resistant disease, relapse or treatment toxicities. Several SNPs are known to be key determinators for inter-individual differences in treatment resistance and toxic side effects. Several studies addressed this hypothesis before, however they concentrated either on single candidate polymorphisms [120, 109, 34], or GWAS approaches without prior knowledge of clinical relevance [134, 133]. Due to the fact that childhood ALL treatment protocols include up to 13 different chemotherapeutic agents, it is hard to evaluate the impact of individual SNPs. In order to progress from the non-targeted GWA studies and limited candidate-gene studies to future clinical trials and further implementation of routine pharmacogenetic screening in childhood ALL, extensive targeted genetic analyses are needed.

In this paper we applied multiplexed targeted sequencing method (described in Paper II) for genotyping of three independent cohorts of childhood ALL, treated on similar protocols. The contribution of inherited genetic variation to treatment response was investigated by associating targeted germline single nucleotide polymorphisms (SNPs) with risk of high minimal residual disease (MRD) levels after remission induction chemotherapy and risk of relapse. We also investigated the effects of several SNPs within the same

gene, protein-protein complex or biological pathway in relation to studied phenotypes. Multiple evidence from associations representing different levels of molecular complexity associate outcome with glucocorticosteroid treatment, immune system functions, signalling pathways and pharmacogenomics of in particular methotrexate and doxorubicin. Results from subsequent analyses were combined together with clinical information in classification and regression tree (CART) analysis, providing a framework for developing an algorithm predicting treatment response based on individual patient SNP profiles.
This study presents first large scale hypothesis-driven screening of thousands of polymorphisms in childhood ALL. The design of the study sets new directions for future pharmacogenetic studies and might serve as a model for other oncology areas. We combined the current state-of-art knowledge of ALL disease mechanisms and pharmacogenetics of the administered drugs curated from a variety of available pharmacogenomics, pathway and protein interaction resources for target selection with a novel, cost-effective method of genotyping by next-generation sequencing. The data analysis in this paper goes beyond analysing the impact of single SNPs, and integrates genotype data across different levels of molecular complexity. This knowledge in the future can be used to help identify patients at risk of relapse at the moment of diagnosis from their individual SNP profile and to adjust their chemotherapy accordingly to prevent treatment failure.

The manuscript presented in this chapter has been submitted to the New England Journal of Medicine and has been formatted according to the journal's specifications. Due to limited amount of text, figures and tables allowed by the journal, additional information was compiled into Supplementary Online Material. Selected figures from the supplementary material are included here after the manuscript.

Extensive targeted SNP profiling predicts early treatment response and risk of relapse in 864 Danish and German childhood ALL patients

Agata Wesołowska-Andersen[1,*], Louise Borst[2,*], Marlene Danner Dalgaard[3], Kirsten Kørup Rasmussen[2], Thomas Sicheritz-Ponten[1], Hans Ole Madsen[4], Hanne Vibeke Marquart[4], Claus R. Bartram[5], Peder Skov Wehner[6], Morten Rasmussen[7], Eske Willerslev[7], Torben Falck Ørntoft[8], Iver Nordentoft[8], Laurent Gautier[1], Søren Brunak[1], Martin Schrappe[9], Martin Stanulla[9], Ramneek Gupta[1] and Kjeld Schmiegelow*[2,10]

[1]Center for Biological Sequence Analysis, The Technical University of Denmark, Copenhagen, Denmark
[2]Department of Paediatrics and Adolescent Medicine, The Juliane Marie Centre, The University Hospital Rigshospitalet, Copenhagen, Denmark
[3]Department of Growth and Reproduction, The University Hospital Rigshospitalet, Copenhagen, Denmark
[4]Department of Clinical Immunology, The University Hospital Rigshospitalet, Copenhagen, Denmark
[5]Institute of Human Genetics, Ruprecht-Karls University, Heidelberg, Germany
[6]Department of Pediatric Hematology and Oncology, H. C. Andersen Children's Hospital, Odense University Hospital, Odense, Denmark
[7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark
[8]Institute for Clinical Medicine, Århus University Hospital, Århus, Denmark
[9]Department for General Pediatrics, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany
[10]Institute of Gynaecology, Obstetrics and Paediatrics, The Faculty of Health Sciences, The University of Copenhagen, Copenhagen, Denmark
*Joint first authorship

ABSTRACT

Background: Despite uniform treatment for childhood acute lymphoblastic leukemia (ALL), considerable interindividual variability exists in treatment response, potentially explained by host genetic variation. Earlier pharmacogenetic studies have concentrated either on single candidate polymorphisms or data-driven genome-wide association studies.

Methods: We applied a novel, cost-effective sequencing-based platform for genotyping thousands of preselected polymorphisms in 864 childhood ALL patients from Nordic NOPHO ALL-92 (n = 143), NOPHO ALL-2000 (n = 232) and German ALL-BFM 2000 (n = 489) cohorts. We explored association of germline single nucleotide polymorphisms (SNPs) to minimal residual disease (MRD) levels after remission induction chemotherapy and relapse risk.

Results: We found 23 and 11 SNPs significantly associated with MRD and relapse, respectively, cross-validated between cohorts. Most significant were rs12546582 (*TNFRSF10A*), rs113708938 (*AKR1C3*) and rs34305100 (*UTS2*) for MRD and rs10502001 (*MMP7*), rs10795242 (*AKR1C3*) and rs28730837 (*MPO*) for relapse risk. Multiple evidence from SNP, gene, protein complex and biological pathway level analyses associate outcome with glucocorticosteroid treatment, immune system functions, G protein-coupled receptor (GPCR) signaling and pharmacogenomics of Methotrexate and Doxorubicin. Categorical regression tree analysis identifies three patient groups with distinct outcome profiles, ranging from 2.5% relapse risk for the best and 77.2% for the worst group, constituting 72% and 7.5% of analyzed patients ($P < 0.001$).

Conclusions: Extensive hypothesis-driven screening of thousands of polymorphisms allows integrative analyses of multiple genotypes grouped by functional modules providing multiple layers of evidence for the hypotheses. This study sets new directions for future pharmacogenetic studies in cancer - moving us closer to the era of personalized medicine.

*Corresponding author: Kjeld Schmiegelow, Department of Pediatrics, The Juliane Marie Centre, The University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. Telephone +45 35451357, Fax +45 35454524, e-mail: kschmiegelow@rh.dk

1. Introduction

The current cure rates for childhood acute lymphoblastic leukemia (ALL) after first-line therapy approach 80-85% in the developed countries [13, 11]. However, even

within risk-adapted treatment groups, there is a wide diversity in the patients' treatment response potentially explainable by genetic variation of both tumor- and host-related factors [11, 4]. One of the strongest, independent predictors of relapse in childhood ALL is high minimal residual disease (MRD) levels at the end of remission induction therapy [13, 11, 4, 2, 10]. Several candidate gene studies demonstrate associations of inherited polymorphisms with post-induction MRD levels and cure rates in childhood ALL [14, 5, 12]. Recent genome-wide association studies (GWAS) identified 102 SNPs associated to MRD levels [18] and 134 SNPs associated to relapse risk [17], however the biological relevance and interpretation of most of those polymorphisms remains unclear. This hinders the translation of such findings to routine pharmacogenetic screening and treatment stratification in childhood ALL. Recently, we reported a novel cost-effective multiplexed targeted sequencing method for screening of approximately 25,000 targeted clinically relevant SNPs in multiple samples simultaneously [16]. This extends the candidate-gene approach to thousands of polymorphisms, focusing a genome-wide approach on the important biological domains. We here apply this method in screening of all Danish childhood ALL patients treated according to the NOPHO ALL-1992 and NOPHO ALL-2000 protocols to explore the genetic components of early treatment response (MRD) and relapse risk, and cross-validating results in ALL-BFM 2000 cohort. Beyond conventional single SNP associations, we present integrative analyses of multiple SNPs acting in the same genes, protein-protein complexes or pathways and define the most important biological mechanisms and drugs influencing the phenotype.

2. Materials and Methods

We genotyped 864 patients from three childhood ALL cohorts (Table 1) for a panel of potentially clinically relevant SNPs by means of multiplexed targeted sequencing as described previously [16]. The genotyping panel consisted of approximately 25,000 SNPs applied for the NOPHO ALL-2000 cohort, later updated to 34,000 SNPs for genotyping of the NOPHO ALL-92 and ALL-BFM 2000 cohorts (Supplementary Figure 1). The selection of genes and polymorphisms involved those previously associated with treatment response, genes influencing pharmacogenomics of drugs administered in ALL treatment, and genes involved in additional aspects of response to chemotherapy including immune system, apoptosis or DNA repair genes. The list was then expanded with first-order protein-protein interactions to a total of 2,349 investigated genes with SNPs residing in their coding and regulatory regions selected for assay. The remission induction therapy in the ALL-BFM 2000 protocol involves a combination of four drugs (Vincristine, Doxorubicin, glucocorticosteroid, L-Asparaginase), while in the NOPHO protocol three drugs are used (no L-Asparaginase) with 30% lower doses of Doxorubicin. Accordingly, the MRD levels are lower in the BFM cohort. To allow integrative association analyses of polymorphisms with MRD levels across ALL cohorts, we used as cut-off the approximate median value for each cohort (MRDmedian NOPHO = 10-3, MRDmedian BFM = 10-4), resulting in similar distribution of patients into high and low MRD groups. Due to differences in cohort sizes, treatment protocol, and measurement methods, we considered SNPs associated to MRD levels if they reached *P*-values below 0.05 in the BFM cohort and if observed allele frequencies and odds ratios in the NOPHO cohort were consistent with those observed in the BFM cohort. Since occurrence of relapse is a more precisely defined phenotype, we required *P*-values below 0.05 in both cohorts to consider the SNPs associated. Besides investigating contribution of individual variations to phenotype, the effects of multiple SNPs acting in the same gene, protein-protein complex or biological pathway were investigated using the CMAT [19] test and neural network models, and subsequently combined with clinical information in classification and regression tree (CART) analysis. Detailed methods are outlined in the Supplementary Online Material (SOM).

3. Results

A total of 864 childhood ALL patients from three cohorts were genotyped for between 25,000 to 34,000 SNPs by size-controlled DNA fragmentation, sample-specific barcoding, sample pooling before SNP-targeted DNA fragment capture, and sequencing. Of the genotyped patients, 820 fulfilled the quality control (QC) criteria and European ancestry requirement (Supplementary Figure 2 & 3). Majority of genotyped variants had very low minor allele frequencies (MAF) and were excluded from the single SNP analysis. Finally, 3,274 SNPs fulfilling the QC were tested for association with post-induction MRD status in 172 NOPHO patients and 4,101 SNPs in 420 BFM patients, respectively. The relapse risk was investigated using 4,260 SNPs in the NOPHO cohorts and 3,865 SNPs in the BFM cohort (in 351 and 424 patients, respectively).

3.1 Single SNP associations

The Manhattan plots for both analyses demonstrate several loci associated with treatment response (Figure 1). The associated QQ plots show good agreement with the

null distribution and absence of genomic inflation (Supplementary Figure 5). After correcting for multiple testing using up to one million adaptive permutations [1], a total of 88 and 82 SNPs were associated with MRD levels in the NOPHO and BFM cohorts, respectively, and 188 and 152 SNPs were associated to relapse risk with *p*-values below 0.05 (data not shown). The 23 SNPs for MRD levels and 11 SNPs for relapse risk, cross-validated between cohorts, are listed in Table 2 and 3. For the replicated relapse-associated SNPs Kaplan-Meier survival analysis was performed showing a general tendency of gene dose effects with the *P*-values for the log-rank trend test ranging from 4.76*10-6 for rs3216144 (*MMP7*) to 0.025 for rs35721373 (*DYSF*) (Supplementary Figure 6).

3.2 Multiple SNP associations

Subsequently we investigated the effects of multiple coding SNPs residing in the same genes by testing for accumulation of variants in the 446 and 388 genes with at least two sequenced variants in association with MRD level and relapse risk, respectively. In MRD analysis 12 genes achieved *P*-values lower than 0.05 after one million permutations in the CMAT test. After manual inspection of the contributing variations, excluding results driven by single SNPs, two of the genes presented meaningful associations contributed by five polymorphisms in the gene *SLCO1A2* and nine in *ABCB5* (Table 4). In relapse risk analysis 19 genes achieved *P*-values lower than 0.05, and after manual inspection of the contributing variations only gene *AKR1C4* presented meaningful association contributed by three SNPs (Supplementary Table 1).
Similarly we tested the effects of multiple coding SNPs in genes forming a protein-protein complex. A total of 196 and 170 virtual pull-down protein-protein complexes were assessed for association to MRD levels and relapse risk respectively, resulting in 11 and 6 associated complexes with permutation *P*-values below 0.05. After manual inspection of the contributing variations, two complexes presented meaningful associations contributed by multiple SNPs in different genes to MRD levels: a complex of *CASP7* and *MVK* ($P = 0.001$) and a complex comprising of *SLCO1B1*, *ABCB11* and *ABCB3* ($P = 0.014$), but none for risk of relapse (Supplementary Table 2).
Finally, we grouped functional SNPs by the biological pathways and assessed the relevance of the pathways to phenotypes by training artificial neural networks on different combinations of SNPs from each pathway, allowing non-linear correlations between SNPs (details in SOM). The top associated Reactome [3] pathways ranked by Matthew's correlation coefficient [9] (MCC, ranging from 0 to 1) were 'Immune system' (MCC = 0.42, AUC = 0.73) for MRD and 'Metabolism of lipids and lipoproteins' (MCC = 0.40, AUC = 0.62) for relapse (Supplementary Table 3 and 4). Similarly, we investigated 11 drug metabolism pathways from PharmGKB [8] relevant for administered drugs, and the top pathways for MRD analysis were the 'Thiopurine pathway' (MCC = 0.34, AUC = 0.65), 'Glucocorticoid Pathway Transcription Regulation, Pharmacodynamics' (MCC = 0.34, AUC = 0.66) and 'Methotrexate Pathway, Pharmacokinetics' (MCC = 0.33, AUC = 0.67), while for relapse risk 'Glucocorticoid Pathway Transcription Regulation, Pharmacodynamics' (MCC = 0.28, AUC = 0.62), 'Methotrexate Pathway, Pharmacokinetics' (MCC = 0.28, AUC = 0.63) and 'Doxorubicin Pathway, Pharmacokinetics' (MCC = 0.27, AUC = 0.63) (Supplementary Table 5).

3.3 CART analyses

To evaluate and illustrate the combined significance of the findings with respect to relapse risk, CART analysis was applied to sequentially sub-classify the cohort in a multivariate modeling, including genotypes of 11 replicated SNPs, number of variants in the *AKR1C4* gene, pathway profiles for top ten Reactome pathways and all investigated PharmGKB pathways and clinical information on MRD levels, white blood cell count, cytogenetics, age and sex (Figure 2). Post-induction MRD levels and white blood cell count (WBC) were the strongest predictors of outcome, and combined with the genomic background information we identified three groups of patients significantly ($P < 0.001$) differing in their risk of relapse. Especially for the group of patients with low MRD levels pathway profiles improved outcome prediction. By this approach we identified two extreme subgroups of patients with a 6-years cumulative risk of relapse of 2.5% (CI95%: 0.5 4.5%) for the best outcome group (72% of all analyzed patients with available MRD) and 77.2% (CI95%: 50.2 89.5%) in the worst outcome group (7.5% of all analyzed patients).

4. Discussion

This study of three independent ALL cohorts using cost-effective targeted genotyping and front-line bioinformatic analyses provides a novel biology- and pharmacology-driven approach for outcome prediction. Treatment is the strongest predictor of relapse in childhood ALL, and the selection of SNPs based on current knowledge of pharmacogenomics, disease mechanisms, and protein interactions allowed us to explore the complete landscape of known relevant pharmacogenomic domains together with their interactome.
The application of multiplexed targeted sequencing for

genotyping significantly reduced costs and provided flexibility to meet the needs of this hypothesis-driven study. Such a large-scale candidate gene and SNP panel facilitates not only single SNP investigations, but also associations of multiple variations grouped by their putative function. This allowed us to test for overrepresentation of variants within genes or protein-protein complexes including rare variants, which are normally excluded from association analyses, but have been shown to accumulate in disease-related genes [19]. Additionally, associations of combinations of SNPs grouped by biological pathways are tested with neural network models, enabling detection of meaningful non-linear SNP interactions by reducing the search space. The results obtained through those strategies provide conclusions at different levels of genomic complexity, collectively emphasizing the importance of the same biological mechanisms for the phenotype.

Most of the 23 SNPs associated to post-induction MRD levels reside in genes involved in metabolism and transport of steroids, Doxorubicin or Methotrexate, or interact with those drugs (as detailed in SOM). Variant accumulation test indicated importance of drug transporters: *SLCO1A2* involved in transport of steroid compounds and *ABCB5* in transport of doxorubicin, while the same test performed on protein-protein complexes supported the aforementioned findings by identifying two complexes that included genes involved in the steroid transport and metabolism: *ABCB3*, *SLCO1B1* and *ABCB11* in the first complex and *MVK* and *CASP7* in the second. Lastly, the Reactome pathways important for MRD levels represented immune system functions, signaling by G protein-coupled receptor (GPCR), hemostasis, and transmembrane transport of small molecules. The last pathway confirms the role of pharmacogenetics in drug response including mostly members of the ATP-binding cassette (ABC) transporters or soluble ligand carrier (SLC) families involved in drug influx and efflux, while GPCR signaling pathways emphasize further the importance of glucocorticosteroid therapy as the action of those drugs is triggered by the membrane-associated GPCRs and activation of their downstream signaling cascades [15]. Accordingly, these data provide important biological input to the many clinical studies of glucocorticosteroid therapy [6, 7], and give credit to the present targeted SNP exploration approach. Similarly, analysis of drug metabolic pathways from PharmGKB indicated the metabolism of thiopurines (or endogenous purines), pharmacodynamics of glucocorticosteroids and pharmacokinetics of Methotrexate (or folates) to be most predictive of the outcome. Clearly, transport, metabolism and downstream effects of glucocorticosteroid therapy comprise a consistent theme for all the associations.

Prediction of relapse risk resulted in 11 cross-validated SNPs residing in genes previously suggested as markers for leukemia aggressiveness, involved in steroid response or implicated in resistance mechanisms or toxicities of certain drugs (addressed in more detail in SOM). No rare variant accumulation was observed for relapse analysis potentially due to the relatively small numbers of affected patients. However, we observed three non-synonymous coding SNPs with MAF higher in relapse patients in the steroid metabolism gene *AKR1C4*. Similarly, pathway analysis revealed importance of the GPCR signaling, homeostasis and immune system, as well as metabolism of lipids and lipoproteins and cell cycle. The top associated pathway 'Metabolism of lipids and lipoproteins' contained genes involved in steroid metabolism, pointing again to importance of such drugs for outcome. Additionally, most of the SNPs pointing to importance of cell cycle pathway reside in genes related to methotrexate and doxorubicin. Not surprisingly, the three top drug metabolism pathways for relapse risk corresponded to the pharmacodynamics of glucocorticoids and pharmacokinetics of Methotrexate and Doxorubicin. Undoubtedly, this reflects the precedence of glucocorticoid therapy, but also emphasizes the importance of the cell cycle arresting agents like Methotrexate and Doxorubicin.

As risk factors linked to host genomics, leukemia biology, treatment response markers, and drug pharmacokinetics pile up, CART analysis allows integration of the different layers of molecular complexity with patients' clinical characteristics to identify groups of patients with distinct treatment outcomes and accordingly candidates for different treatment approaches. Among the identified three groups of patients with a highly significant difference in relapse risk, the group with a 2.5% projected risk of relapse comprised 72% of analyzed patients, whereas the group with a 77.2% risk of relapse included 45% of included relapses. Not surprisingly, combining patients' genotypes with their clinical features can explain the treatment outcome better than single SNPs. Further, in contrast to conventional multivariate Cox regression analyses, CART approach yields classifications easier to perceive and apply in the clinical setting. If additional independent childhood ALL studies confirm these associations, these findings are of sufficient power to be considered in future treatment stratification.

Compared to other large-scale studies, this study has a clear advantage of specifically targeting potential causative variants, without need to investigate genomic LD patterns or conduct follow-up fine-mapping studies. Functional prioritization of genotyped SNPs eliminates the difficulty of mapping variants to genes and the genotypes can be directly combined in functional analyses. Recently Yang *et al.* published two papers on GWAS analysis of childhood ALL patients with respect to MRD levels after remission induction therapy [18] and risk of relapse [17]. All the associated SNPs from the first study were included in this analysis, however few of the SNPs were found associated in either the NOPHO or BFM cohorts (Supplementary Table 6), Only 1 of the 134 SNPs associated to relapse risk was genotyped in this study, as all associated SNPs

resided in non-coding regions, but it was not significant in this study. We were not able to assess associations of the other SNPs. However, lack of replication of MRD associated SNPs suggests that the results potentially could reflect the applied treatment protocol or ethnicity of investigated patients. Moreover, since the associated SNPs in the studies of Yang and coworkers in most cases were difficult to map to the gene on which they exert their effects, it is difficult to elucidate the involved biological mechanisms.

Our study combines large-scale genetic investigations of a GWAS approach with the biological relevance of a candidate-gene approach. The major strength of our approach is that a panel of functional SNPs was selected with various prior assumptions of relevance for childhood ALL. Therefore, the results obtained through the method presented here are easier to interpret, and thus implement in the clinic based on existing knowledge of ALL disease mechanisms and pharmacogenomics of administered drugs. Not covering the scope of the whole genome potentially leads to overlooking unknown important genetic components of treatment response, and such exploration will require genome-wide screening approaches such as GWAS or whole exome/transcriptome/genome sequencing, although the latter is still burdened by its relatively high costs, and required patient numbers for statistical significance.

In conclusion, this study introduces new potential targets for future pharmacogenetic studies in childhood ALL. In addition to potentially allowing future risk grouping based on individual SNP profiles, using a similar approach for the multiple toxicities burdening childhood ALL patients could pave the way for higher cure rates with less toxicity.

Conflict of interest

The authors declare no conflict of interest.

Author Contribution

KS and RG conceived of and headed the project, AWA and LB wrote the manuscript, AWA performed the presented analyses and compiled the work, LB, MDD and KKR performed the laboratory work, TSP and RG assisted with the neural network analysis, HOM, HVM, and CRB provided the MRD measurements, MS and MS provided samples from BFM cohort, PSW provided samples from western Denmark, MR, EW, TØ and IN provided the sequencing, LG provided technology consultation and laboratory resources, SB provided computational infrastructure. All authors provided critical input to the project and manuscript and approved the final manuscript.

Acknowledgements

This study was supported by grants from The Danish Cancer Society (Grant numbers R2-A56-09-S2 and R20-A1156-10-S2), The Danish Childhood Cancer Foundation, The Ministry of Health (Grant number 2006-12103-250), The Novo Nordisk Foundation and The Danish Research Council for Health and Disease (Grant numbers 271-06-0278, 271-08-0684). Kjeld Schmiegelow holds the Danish Childhood Cancer Foundation Professorship in Pediatric Oncology. Ramneek Gupta is supported by a grant from the Danish National Research Foundation to the Sino-Danish Breast Cancer Research Center. The study was conducted in the framework of the International BFM Study Group.

References

[1] J. Besag and P. Clifford, "Sequential monte carlo p-values", *Biometrika*, Vol. 78, No. 2, pp. 301–304, 1991.

[2] D. Campana, "Minimal residual disease in acute lymphoblastic leukemia", *ASH Education Program Book*, Vol. 2010, No. 1, pp. 7–12, 2010.

[3] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, et al., "Reactome: a database of reactions, pathways and biological processes", *Nucleic acids research*, Vol. 39, No. suppl 1, pp. D691–D697, 2011.

[4] M. Davidsen, K. Dalhoff, and K. Schmiegelow, "Pharmacogenetics influence treatment efficacy in childhood acute lymphoblastic leukemia", *Journal of pediatric hematology/oncology*, Vol. 30, No. 11, pp. 831–849, 2008.

[5] S. Davies, M. Borowitz, G. Rosner, K. Ritz, M. Devidas, N. Winick, P. Martin, P. Bowman, J. Elliott, C. Willman, et al., "Pharmacogenetics of minimal residual disease response in children with B-precursor acute lymphoblastic leukemia: a report from the Children's Oncology Group", *Blood*, Vol. 111, No. 6, pp. 2984–2990, 2008.

[6] P. Gaynon and A. Carrel, "Glucocorticosteroid therapy in childhood acute lymphoblastic leukemia.", *Advances in experimental medicine and biology*, Vol. 457, p. 593, 1999.

[7] P. Gaynon, R. Lustig, et al., "The use of glucocorticoids in acute lymphoblastic leukemia of childhood. Molecular, cellular, and clinical considerations.", *Journal of pediatric hematology/oncology*, Vol. 17, No. 1, p. 1, 1995.

[8] M. Hewett, D. Oliver, D. Rubin, K. Easton, J. Stuart, R. Altman, and T. Klein, "PharmGKB: the pharmacogenetics knowledge base", *Nucleic acids research*, Vol. 30, No. 1, pp. 163–165, 2002.

[9] B. Matthews et al., "Comparison of the predicted and observed secondary structure of T4 phage lysozyme.", *Biochimica et biophysica acta*, Vol. 405, No. 2, p. 442, 1975.

[10] C. Nyvold, H. Madsen, L. Ryder, J. Seyfarth, A. Svejgaard, N. Clausen, F. Wesenberg, O. Jonsson, E. Forestier, and K. Schmiegelow, "Precise quantification of minimal residual disease at day 29 allows identification of children with acute lymphoblastic leukemia and an excellent outcome", *Blood*, Vol. 99, No. 4, pp. 1253–1258, 2002.

[11] C. Pui, D. Pei, D. Campana, W. Bowman, J. Sandlund, S. Kaste, R. Ribeiro, J. Rubnitz, E. Coustan-Smith, S. Jeha, et al., "Improved prognosis for older adolescents with acute lymphoblastic leukemia", *Journal of Clinical Oncology*, Vol. 29, No. 4, pp. 386–391, 2011.

[12] J. Rocha, C. Cheng, W. Liu, S. Kishi, S. Das, E. Cook, J. Sandlund, J. Rubnitz, R. Ribeiro, D. Campana, et al., "Pharmacogenetics of outcome in children with acute lymphoblastic leukemia", *Blood*, Vol. 105, No. 12, pp. 4752–4758, 2005.

[13] K. Schmiegelow, E. Forestier, M. Hellebostad, M. Heyman, J. Kristinsson, S. Söderhäll, and M. Taskinen, "Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia", *Leukemia*, Vol. 24, No. 2, pp. 345–354, 2009.

[14] M. Stanulla, E. Schaeffeler, T. Flohr, G. Cario, A. Schrauder, M. Zimmermann, K. Welte, W. Ludwig, C. Bartram, U. Zanger, et al., "Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia", *JAMA: the journal of the American Medical Association*, Vol. 293, No. 12, pp. 1485–1489, 2005.

[15] J. Tasker, S. Di, and R. Malcher-Lopes, "Rapid glucocorticoid signaling via membrane-associated receptors", *Endocrinology*, Vol. 147, No. 12, pp. 5549–5556, 2006.

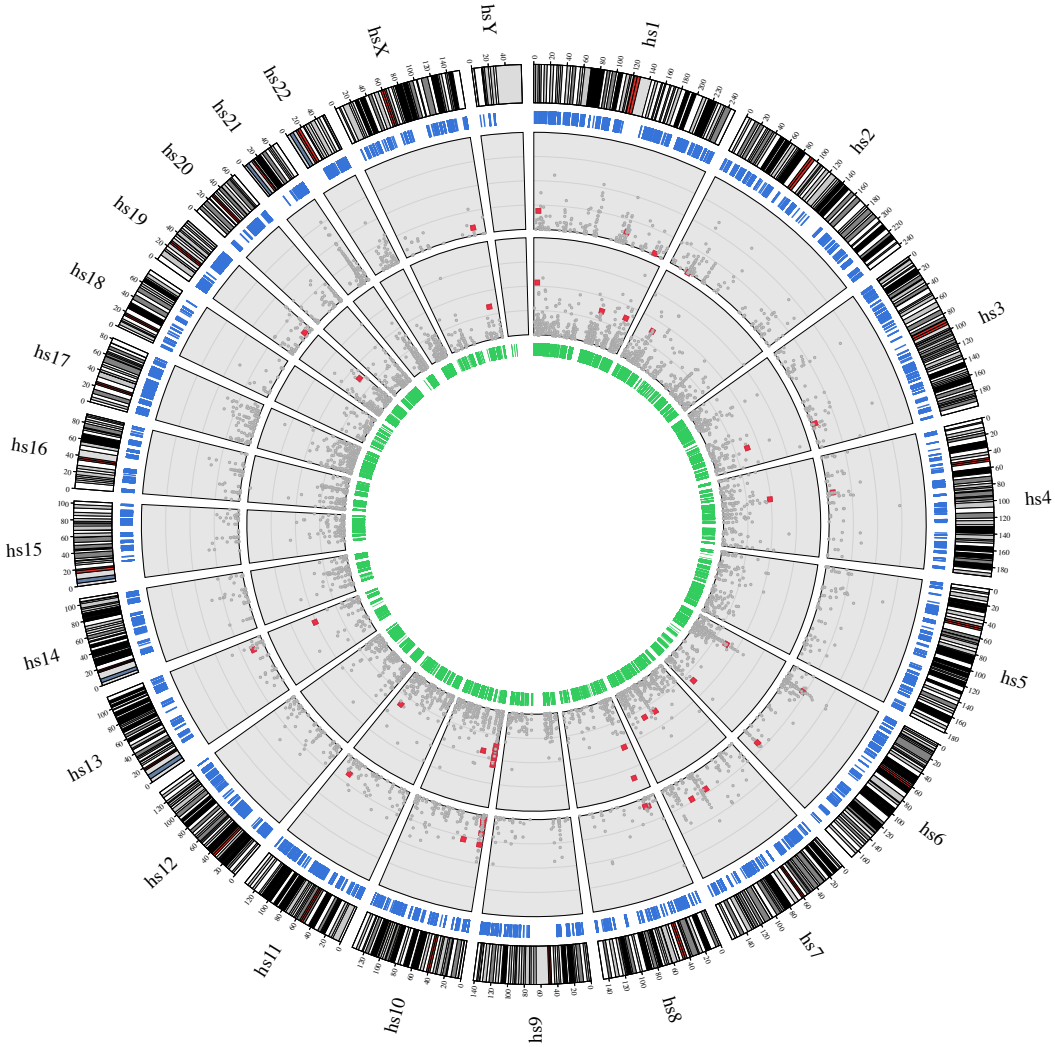
[16] A. Wesolowska, M. Dalgaard, L. Borst, L. Gautier, M. Bak, N. Weinhold, B. Nielsen, L. Helt, K. Audouze, J. Nersting, et al., "Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia", *Leukemia*, Vol. 25, No. 6, pp. 1001–1006, 2011.

[17] J. Yang, C. Cheng, M. Devidas, X. Cao, D. Campana, W. Yang, Y. Fan, G. Neale, N. Cox, P. Scheet, et al., "Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia.", *Blood*, 2012.

[18] J. Yang, C. Cheng, W. Yang, D. Pei, X. Cao, Y. Fan, S. Pounds, G. Neale, L. Treviño, D. French, et al., "Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia", *JAMA: the journal of the American Medical Association*, Vol. 301, No. 4, pp. 393–403, 2009.

[19] M. Zawistowski, S. Gopalakrishnan, J. Ding, Y. Li, S. Grimm, and S. Zöllner, "Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes", *American journal of human genetics*, Vol. 87, No. 5, p. 604, 2010.

Figure 1
A. MRD



B. Relapse

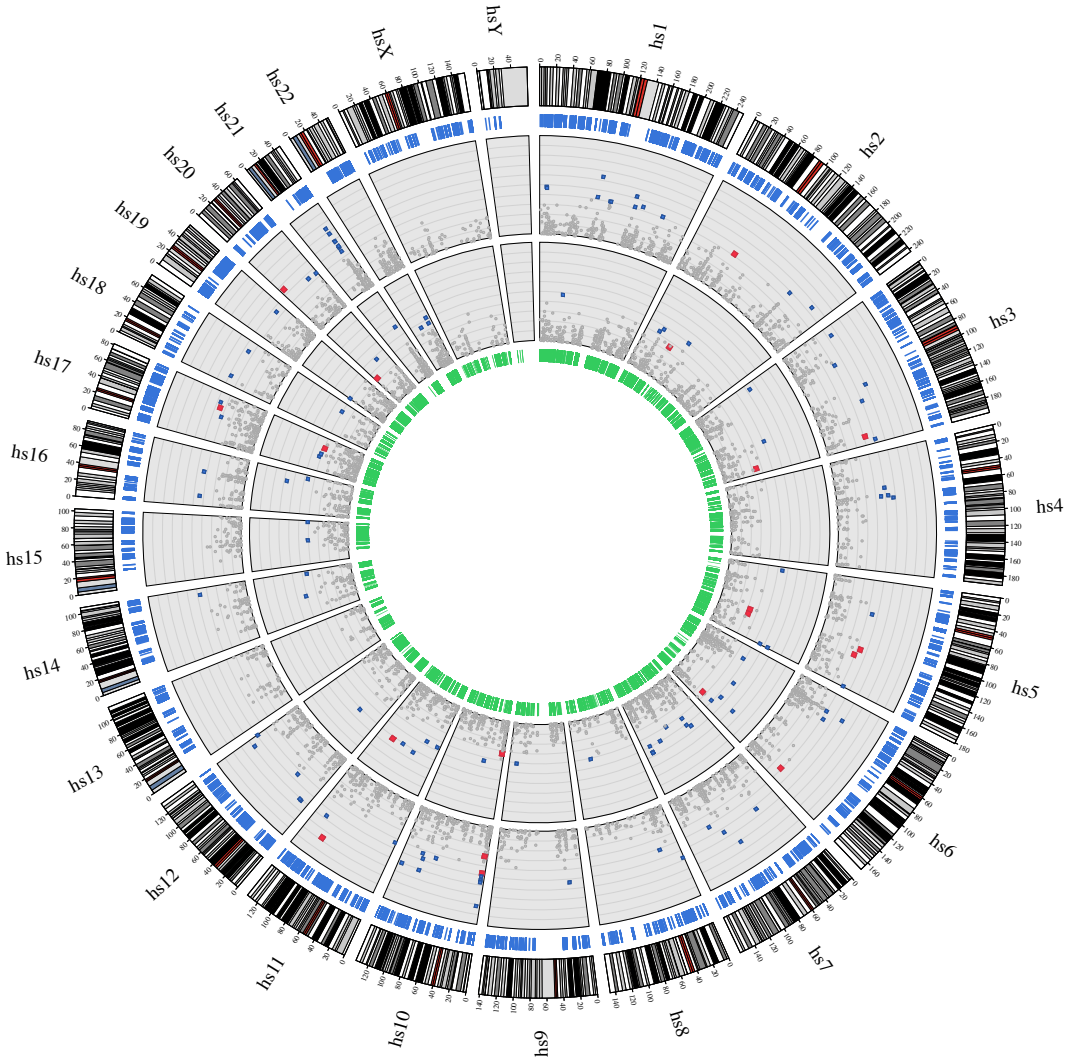
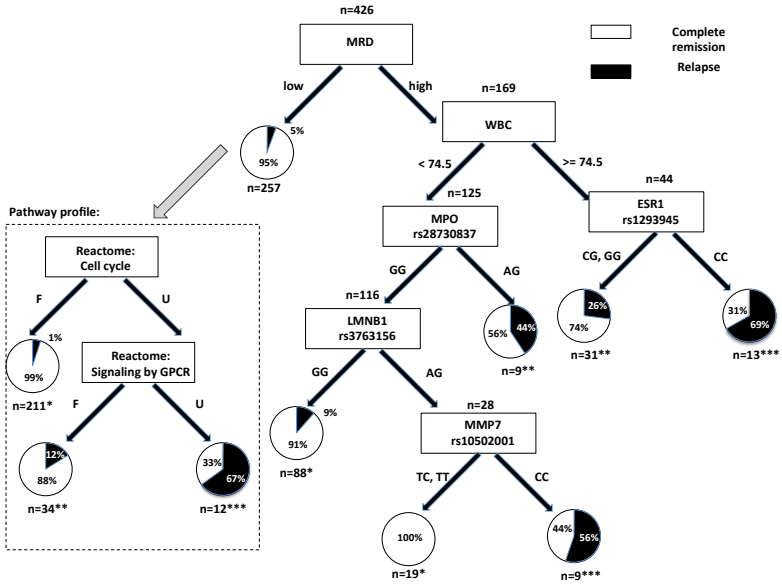

hsY
hs1
hs2
hs3
hs4
hs5
hs6
hs7
hs8
hs9
hs10
hs11
hs12
hs13
hs14
hs15
hs16
hs17
hs18
hs19
hs20
hs21
hs22
hsX

Figure 2

A.


n=426
MRD
Complete remission
Relapse
low
high
n=169
WBC
5%
95%
n=257
< 74.5
>= 74.5
n=125
MPO rs28730837
n=44
ESR1 rs1293945
Pathway profile:
Reactome: Cell cycle
GG
AG
CG, GG
CC
n=116
LMNB1 rs3763156
26%
74%
n=31**
31%
69%
n=13***
F
U
1%
99%
n=211*
Reactome: Signaling by GPCR
56%
44%
n=9**
GG
AG
9%
91%
n=88*
n=28
MMP7 rs10502001
F
U
12%
88%
n=34**
33%
67%
n=12***
TC, TT
CC
100%
n=19*
44%
56%
n=9***

B.

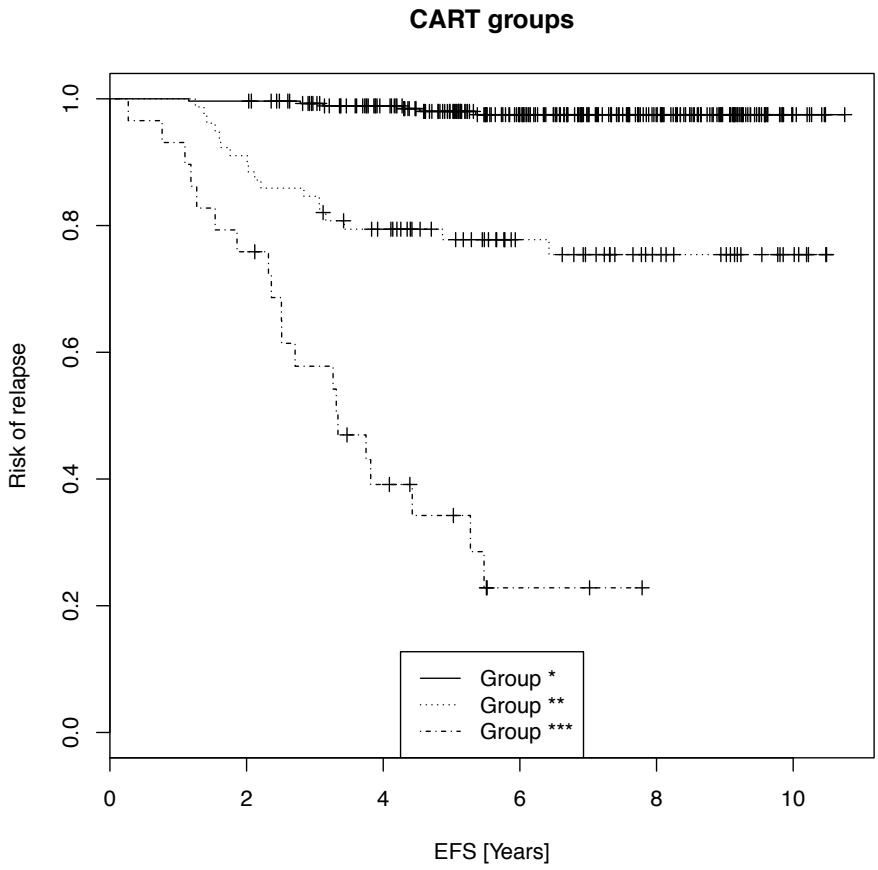

CART groups
Risk of relapse
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
EFS [Years]
Group *
Group **
Group ***

Table 1. Characteristics of patients included in the study from three cohorts: NOPHO ALL-92, NOPHO ALL-2000 and BFM ALL 2000. Patients failing QC steps are excluded from all the analysis. Patients included in MRD analysis represent all the patients for whom MRD measurements were available, in the BFM cohort additionally patients treated with dexamethasone during induction therapy were excluded. For relapse analysis patients with events other than relapse were excluded. MRD = minimal residual disease, WBC = white blood cell count, BCP-ALL = B-cell precursor ALL, T-ALL = T-cell ALL, SR = standard risk, IR = intermediate risk, HR = high risk

	NOPHO ALL-92			NOPHO ALL-2000			BFM ALL 2000		
	Genotyped N = 143	**MRD analysis N = 0**	**Relapse analysis N = 131**	**Genotyped N = 232**	**MRD analysis N = 172**	**Relapse analysis N = 221**	**Genotyped N = 489**	**MRD analysis N = 264**	**Relapse analysis N = 435**
Gender									
Male (%)	91 (63.6)	NA	86 (65.6)	129 (55.6)	95 (55.2)	120 (54.3)	294 (60.1)	165 (61.6)	255 (58.6)
Female (%)	52 (36.4)		45 (34.4)	103 (44.4)	77 (44.8)	101 (45.7)	195 (39.9)	103 (38.4)	180 (41.4)
Age (median in years)	4.24 (1.13 – 14.57)	NA	4.18 (1.13 – 14.57)	4.48 (1.31 – 14.97)	4.51 (1.31 – 14.97)	4.48 (1.31 – 14.97)	5.28 (1.01 – 17.95)	4.43 (1.01 – 14.98)	4.83 (1.01 – 14.98)
WBC (median x 10^9/L	8.3 (0.5 – 815)	NA	8.3 (0.5 – 815)	10.4 (0.5 – 604)	10.35 (0.9 – 604)	10.2 (0.5 – 604)	16.8 (0.3 – 751)	13.9 (0.3 – 751)	15.5 (0.3 – 751)
Immunophenotype									
BCP-ALL (%)	127 (88.8)	NA	116 (88.5)	197 (84.9)	151 (87.8)	191 (86.4)	404 (82.6)	233 (86.9)	373 (85.7)
T-ALL (%)	16 (11.2)		15 (11.5)	35 (15.1)	21 (12.2)	30 (13.6)	83 (17)	34 (12.7)	60 (13.8)
Unknown (%)	0 (0)		0 (0)	0 (0)	0 (0)	0 (0)	2 (0.4)	1 (0.4)	2 (0.5)
Risk group									
SR (%)	58 (40.5)	NA	53 (40.5)	89 (38.4)	68 (39.5)	86 (39)	180 (36.8)	111 (41.4)	168 (38.6)
IR (%)	55 (38.5)		51 (38.9)	65 (28)	50 (29.1)	62 (28)	224 (45.8)	121 (45.1)	204 (46.9)
HR (%)	30 (21)		27 (20.6)	76 (32.7)	52 (30.2)	71 (32.1)	85 (17.4)	36 (13.4)	63 (14.5)
Other (%)	0 (0)		0 (0)	2 (0.9)	2 (1.2)	2 (0.9)	0 (0)	0 (0)	0 (0)
Events									
Induction failure (%)	0 (0)	NA	0 (0)	1 (0.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Resistant disease (%)	2 (1.4)		0(0)	6 (2.6)	5 (2.8)	0 (0)	0 (0)	0 (0)	0 (0)
Relapse (%)	23 (16.1)		22 (16)	24 (10.3)	18 (10.5)	24 (10.8)	51 (10.4)	30 (11.2)	45 (10.3)
Death in remission (%)	1 (0.7)		0 (0)	1 (0.4)	0 (0)	0 (0)	13 (2.6)	3 (1.2)	0 (0)
Secondary malignancy (%)	1 (0.7)		0 (0)	2 (0.9)	2 (1.2)	0 (0)	7 (1.4)	2 (0.7)	0 (0)
Total events (%)	27 (18.8)		22 (16.8)	34 (14.7)	25 (14.5)	24 (10.8)	71 (14.5)	35 (13)	45 (10.3)
Complete remission (%)	116 (81.2)	NA	109 (83.2)	198 (85.3)	147 (85.5)	197 (89.2)	418 (85.5)	233 (87)	390 (89.7)

Table 2. SNPs associated with MRD levels. SNPs identified in BFM ALL 2000 cohort with P-values below 0.05 and showing similar allele frequencies and odds ratios in NOPHO ALL 2000 cohort. SNPs are sorted by their respective gene names. Cons = consequence of variant on its transcript, WNCG = Within non-coding gene, NSC = non-synonymous coding, OR = odds ratio, MAF high = minor allele frequency in high MRD group, MAF low = minor allele frequency in low MRD group

SNP			BFM ALL 2000				NOPHO ALL 2000				Combined cohorts			
rsID	Gene	Cons	*P*-value	OR	MAF high	MAF low	*P*-value	OR	MAF high	MAF low	MAF high	MAF low	*P*-value	OR
rs61040122	*ABCB1*	Intronic	0.0438	NA	0.014	0	0.1421	NA	0.025	0	0.015	0	0.0293	NA
rs74341718	*ABCC4*	Regulatory	0.009	NA	0.023	0	0.4222	NA	0.012	0	0.015	0	0.1518	NA
rs34395363	*AGT*	Upstream	0.0328	1.96	0.148	0.082	1	1.62	0.087	0.056	0.139	0.078	0.0166	1.92
rs6601899	*AKR1C3*	Upstream	0.0059	1.89	0.191	0.111	0.3846	1.37	0.268	0.210	0.207	0.131	0.0105	1.73
rs113708938	*AKR1C3*	Upstream	0.0109	1.86	0.194	0.114	0.041	2.29	0.241	0.122	0.204	0.116	0.0023	1.96
rs10752002	*AKR1C3*	Upstream	0.0135	1.87	0.189	0.111	0.3235	1.36	0.210	0.163	0.194	0.122	0.0129	1.73
rs10904401	*AKR1C3*	Intronic	0.0151	1.62	0.518	0.399	0.4318	1.32	0.525	0.456	0.519	0.409	0.0063	1.56
rs10904411	*AKR1C3*	Intronic	0.0267	1.76	0.18	0.111	0.1857	2	0.25	0.143	0.189	0.116	0.0147	1.77
rs6601893	*AKR1C3*	Intronic	0.0364	0.64	0.23	0.317	0.35	0.73	0.224	0.283	0.228	0.308	0.0218	0.66
rs1805171	*CFTR*	Splice site	0.0454	1.54	0.257	0.183	0.1475	2.17	0.194	0.1	0.256	0.169	0.0092	1.68
rs162326	*CYP1B1*	WNCG	0.0301	0.60	0.158	0.239	0.8571	0.77	0.159	0.197	0.156	0.231	0.0231	0.61
rs118178942	*ERCC1*	3'UTR	0.0281	NA	0.048	0	0.8	1.72	0.025	0.015	0.039	0.006	0.0814	6.33
rs76253914	*ESR1*	Regulatory	0.0414	NA	0.019	0	0.1111	NA	0.024	0	0.022	0	0.062	NA
rs9332701	*F5*	NSC	0.0394	0.37	0.022	0.059	0.3026	0.28	0.042	0.135	0.024	0.07	0.0138	0.33
rs4339791	*HPRT1*	Intronic	0.0461	1.94	0.189	0.107	0.439	1.54	0.109	0.074	0.168	0.098	0.007	1.865
rs17283597	*MNBL1*	WNCG	0.0299	0.51	0.077	0.140	0.5833	0.76	0.103	0.13	0.082	0.138	0.0232	0.56
rs4986783	*NAT1*	NSC	0.0082	5.27	0.049	0.010	0.5161	NA	0.013	0	0.040	0.007	0.0062	5.45
rs34299487	*NRP1*	Regulatory	0.0187	1.69	0.399	0.281	0.3774	1.81	0.437	0.3	0.407	0.284	0.0093	1.73
rs2163154	*PTS*	Downstream	0.0401	1.52	0.306	0.225	0.1884	1.55	0.284	0.204	0.297	0.22	0.0231	1.50
rs12546582	*TNFRSF10A*	Upstream	0.0004	15.9	0.049	0.0005	0.8571	2.07	0.047	0.023	0.048	0.008	0.0009	6.58
rs139899336	*UGT2A3*	Downstream	0.0111	NA	0.017	0	0.8571	1.73	0.024	0.014	0.018	0.003	0.1631	6.05
rs34305100	*UTS2*	NSC	0.0070	1.82	0.223	0.136	0.1337	1.79	0.25	0.157	0.229	0.141	0.0035	1.81
rs73728204	intergenic	Regulatory	0.0305	0.21	0.009	0.043	0.139	0	0	0.03	0.007	0.04	0.0067	0.17

Table 3. SNPs associated with risk of relapse discovered in both Danish and German cohorts. SNPs are sorted by the *P*-value in combined cohorts. Cons = consequence of the SNP on its transcript, OR = odds ratio, MAF relapse = minor allele frequency in relapse patients, MAF CR = minor allele frequency in complete remission patients, NSC = non-synonymous coding, SC = synonymous coding

SNP			**NOPHO ALL 92 & 2000**				**BFM ALL 2000**				**Combined cohorts**			
rsID	Gene	Cons	***P*-value**	**OR**	**MAF relapse**	**MAF CR**	***P*-value**	**OR**	**MAF relapse**	**MAF CR**	**MAF relapse**	**MAF CR**	***P*-value**	**OR**
rs3216144	*MMP7*	Regulatory	0.0004	0.14	0.045	0.257	0.002	0.32	0.089	0.235	0.074	0.239	6e-06	0.26
rs10502001	*MMP7*	NSC	0.0004	0.14	0.045	0.252	0.002	0.32	0.089	0.235	0.075	0.238	6e-06	0.26
rs10795242	*AKR1C3*	Intronic	0.0086	2.13	0.309	0.173	0.0368	1.85	0.244	0.149	0.272	0.157	0.0005	2.01
rs28730837	*MPO*	NSC	0.047	4.26	0.064	0.016	0.015	3.19	0.058	0.019	0.061	0.017	0.001	3.6
rs6139873	*CHGB*	NSC	0.004	19.56	0.107	0.006	0.039	3.5	0.043	0.013	0.061	0.011	0.0011	5.75
rs1293945	*ESR1*	Regulatory	0.02	1.86	0.588	0.434	0.026	1.95	0.586	0.421	0.587	0.427	0.0013	1.91
rs3763156	*LMNB1*	Intronic	0.0197	3.06	0.19	0.071	0.0396	2.15	0.147	0.074	0.164	0.074	0.0016	2.43
rs55684978	*HTR3D*	SC	0.016	8.08	0.071	0.009	0.032	3.85	0.044	0.012	0.053	0.011	0.0018	4.88
rs1058047	*TMED7*	Splice site	0.0129	8.64	0.083	0.01	0.039	3.2	0.045	0.015	0.056	0.015	0.0031	3.98
rs35721373	*DYSF*	SC	0.015	4.89	0.125	0.028	0.031	2.52	0.086	0.036	0.098	0.035	0.0033	2.96
rs6601899	*AKR1C3*	Intronic	0.0391	1.77	0.306	0.199	0.036	1.8	0.239	0.148	0.267	0.165	0.0033	1.84

Figures

Figure S1. Genomic target of two SNP panels.
The genomic coverage of targeted regions with first and second target capture designs is plotted in blue and green, respectively. Each line represents a 100,000 base pairs, and the maximum value shown here corresponds to 5% coverage in this genomic region.

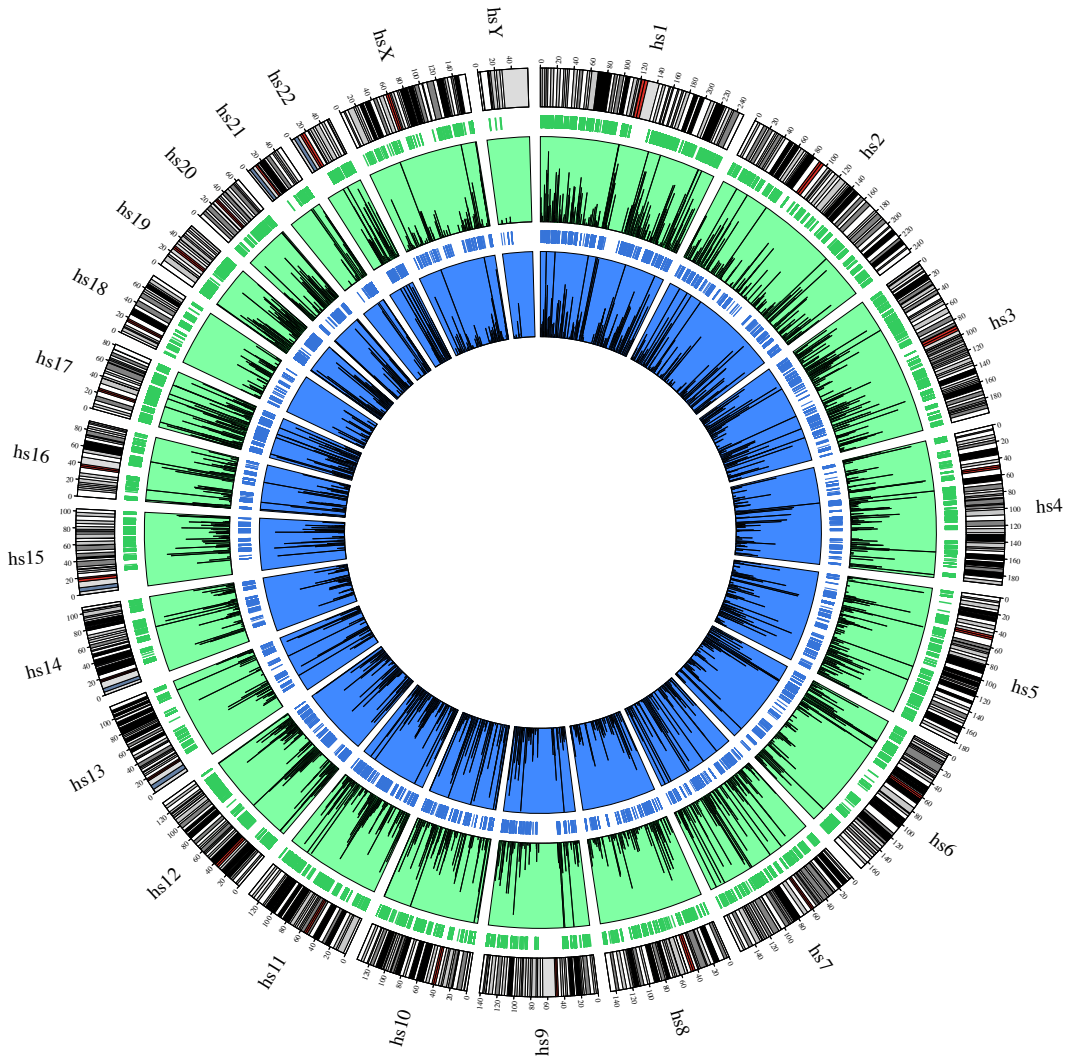


Figure S3. Identity by state for NOPHO ALL-2000 cohort.
In order to investigate the ethnicity of the patients we performed this identity by state analysis. For this purpose 4,295 SNPs genotyped in NOPHO ALL-2000 cohort and overlapping with the HapMap phase 3 data were analyzed by means of principal component analysis. The first two principal components are plotted and the childhood ALL patients are shown in red and the HapMap-CEU population in blue. All the samples, except for one, are grouped together with other individuals of European ancestry. The one sample of Asian origin was excluded from further analysis.

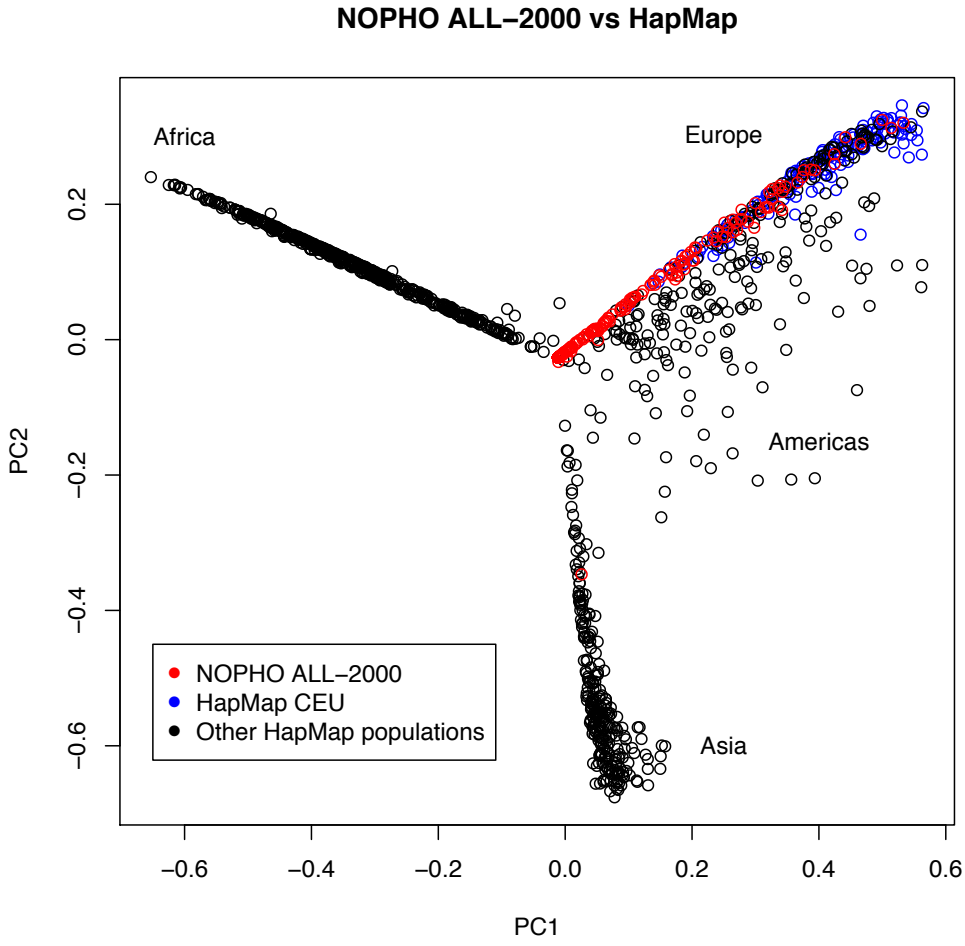


Figure S4. Identity by state for NOPHO ALL-92 and BFM ALL2000 cohorts. In order to investigate the ethnicity of the patients we performed this identity by state analysis. For this purpose approximately 3,862 SNPs genotyped in NOPHO ALL-1992 and German cohorts and overlapping with the HapMap phase 3 data were analyzed by means of principal component analysis. The first two principal components are plotted and the childhood ALL patients are shown in red and the HapMap-CEU population in blue. Majority of the samples are grouped together with other individuals of European ancestry. The one sample of African origin was excluded from further analysis, as well as the 18 samples presenting as outliers in the plot.

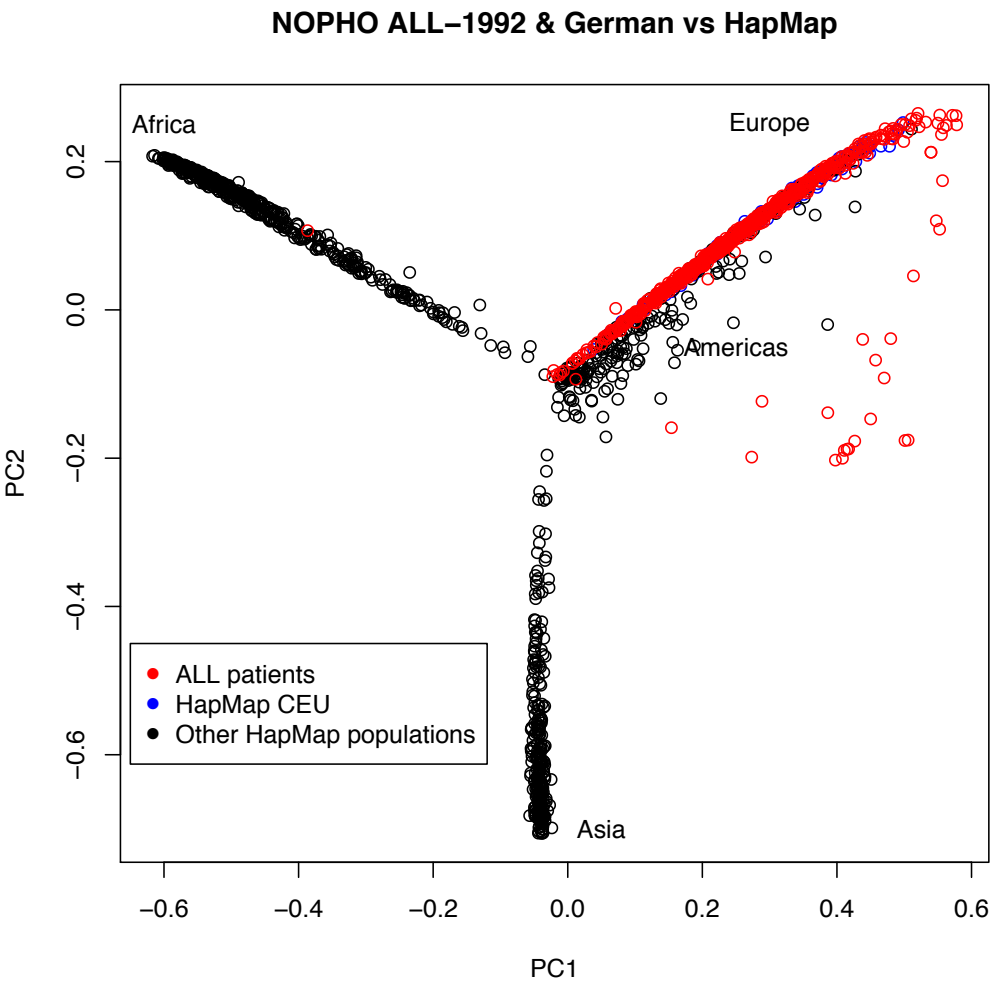


Figure S5. MAF comparison.
Minor allele frequencies (MAF) comparison of the SNPs genotyped in childhood ALL patients with overlapping SNPs genotyped in HapMap and 1000 genomes projects samples. Fisher's exact test was performed for all the SNPs to assess deviation from expected distribution, the SNPs showing slight deviation from that are plotted in blue, while most deviating SNPs are shown in red.

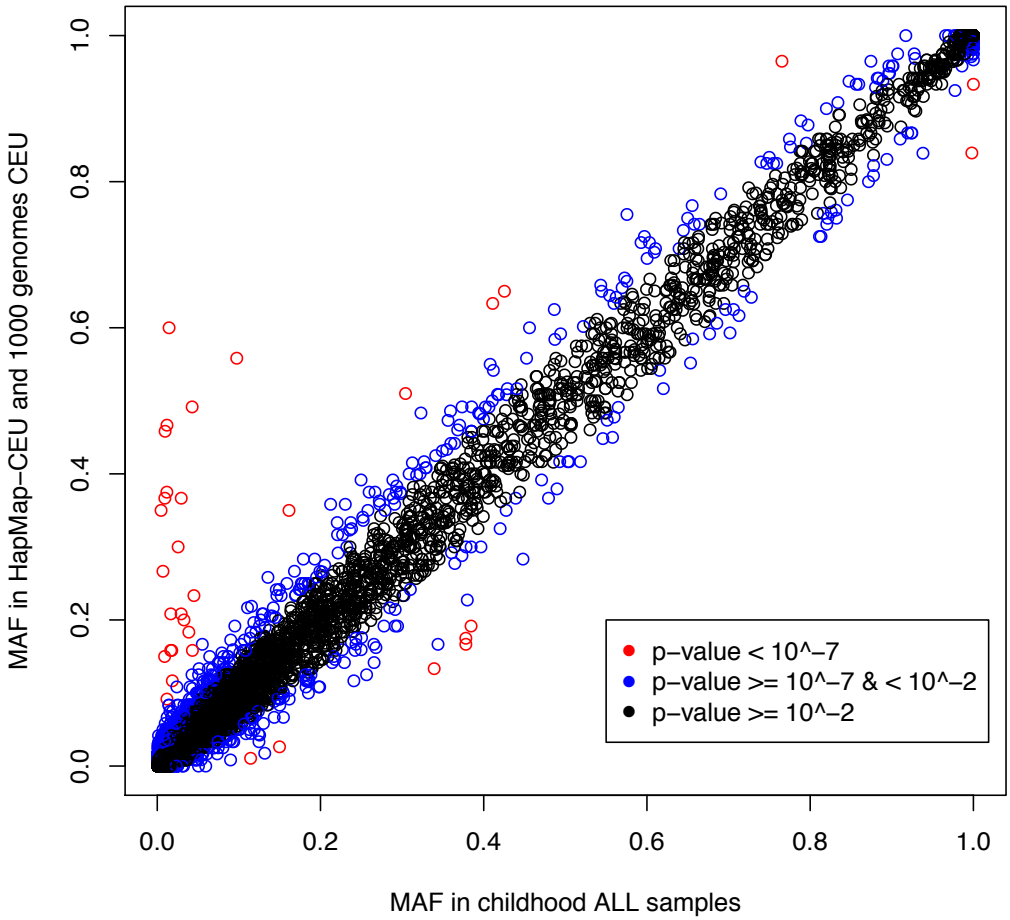
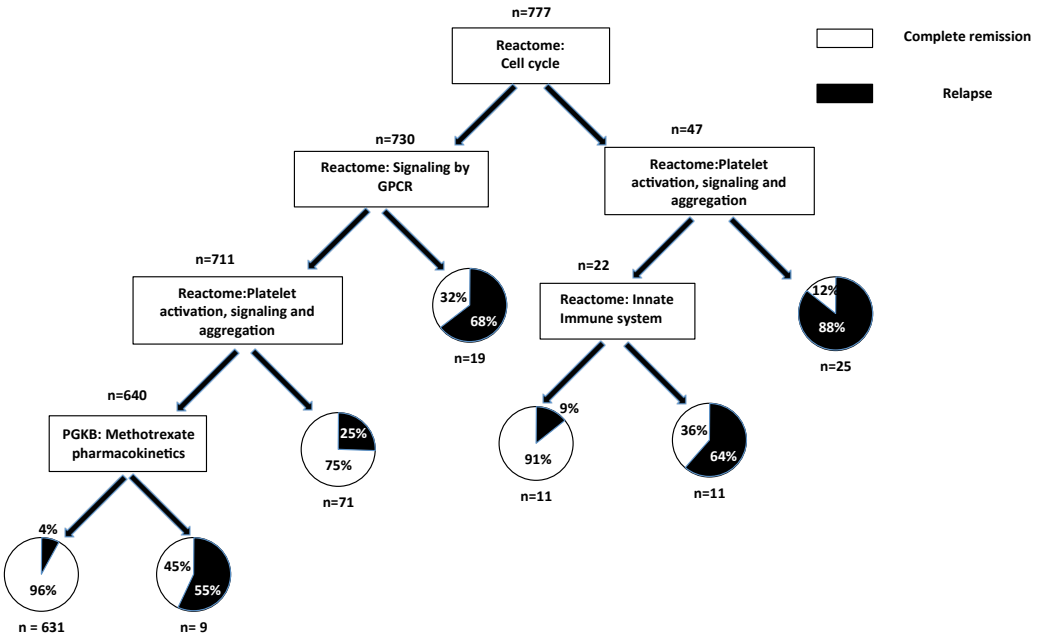


Figure S8. CART analysis – pathways.
CART analysis of relapse risk by pathway profiles from neural network models. This model gives a general idea of involved biological processes, however since each pathway profile consists of up to 15 SNPs, the model is too complex to be used in treatment response predictions. Arrows going to the left represent favourable pathway profiles, while arrows going to the right represent unfavourable pathway profiles.



Part IV

Infections during induction therapy

Chapter 9

Paper IV - Variation in host genetics and infections during induction treatment in childhood acute lymphoblastic leukaemia

Infections remain a significant challenge in treatment of childhood ALL sometimes causing treatment related death [75]. Since relapse rates are decreasing due to more risk adapted and, for some groups, more toxic therapy, focus on toxic events is increasingly essential. Among patients dying from toxicity, infections are the most common cause in which a majority happen during the first months of treatment. Inherited genetic variants may influence the susceptibility to infections during treatment. In this paper we describe the factors influencing the risk of infections and the patients outcome in infectious events (Figure I, Paper IV). We then focus on studying patients' genetic background to identify the risk profiles.
We used the multiplexed targeted sequencing method (described in Paper II) for genotyping of 69 patients for a set of approximately 34,000 genomic variations selected based on prior knowledge of ALL disease mechanisms, pharmacogenetic of administered drugs and immune system functions (SNP selection is described in more detail in Chapter 6.1).
The 69 patients included in the study were treated according to NOPHO ALL-92 protocol at a single centre in Denmark and all the infectious events were registered during the first 7 weeks of treatment constituting the remission induction therapy. The phenotypes used in the study are occurrence of at least one infectious event during induction therapy and positive blood

culture. The investigated cohort is very small and single SNP association analysis yields many results probably burdened by many false positive findings. However, by integrating SNP effects in well defined biological pathways we were able to define the most important mechanisms of susceptibility to infectious events and classify patients into two risk groups with significantly different risk of infections. Certainly, in order to be translated into clinical settings, these findings have to be validated in a larger cohort of patients, however the study demonstrates that host genomic risk profiling is possible even in relatively small sample sizes. The results of this study indicate that susceptibility to infections is a well-defined phenotype with considerate genetic component and therefore genetic screening provides a promising way of identifying patients at risk and adapting individualised supportive care and treatment protocols accordingly.

Variation in host genetics and infections during induction treatment in childhood acute lymphoblastic leukemia

Bendik Lund[1,2,*], Agata Wesolowska-Andersen[3,*], Birgitte Lausen[4], Louise Borst[5], Kirsten Kørup Rasmussen[5], Klaus Muller[5], Helge Klungland[5], Ramneek Gupta[3] and Kjeld Schmiegelow[*5,6]

[1]Dept. of Pediatrics, St Olavs Hospital, Trondheim, Norway
[2]Department for laboratory Medicine, Children's and Women's Health, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway
[3]Center for Biological Sequence Analysis, The Technical University of Denmark, Copenhagen, Denmark
[4]Department of Pediatrics, Rigshospitalet, Copenhagen, Denmark
[5]Department of Paediatrics and Adolescent Medicine, The University Hospital Rigshospitalet, Copenhagen, Denmark
[*]Joint first authorship

ABSTRACT

Purpose.To investigate host genomic variation and its association with risk of infections during induction treatment of children with acute lymphoblastic leukemia.

Patients and Methods. A total of 69 patients aged 1 to 15 years treated at the University Hospital Rigshospitalet, Denmark, from 1992 to 2000 were included. Approximately 34 000 candidate single nucleotide polymorphisms (SNPs) related primarily to pharmacogenomics and immune function were analysed for each patient and associations with infectious events during induction treatment were explored by testing individual SNPs and multiple SNPs grouped by biological pathways.

Results. Forty-eight (70%) of the patients experienced an infectious event, and of these 23 (33% of total) also had a positive blood culture. A total of 103 and 94 SNPs were associated with having an infectious event or having a positive blood culture, respectively. CART analysis demonstrated rs11033797 (*OR51F1*), rs2835265 (*CBR1*), rs28627172 (*POLDIP3*) and rs1129844/CM072923 (*CCL11*) to be highly predictive of outcome characterizing 40 of 45 (89%) infectious event-patients with 98% accuracy. Pathway analysis identified variations in GPCR downstream signalling, Bile acid and bile salt metabolism (both involved in glucocorticosteroid pharmacokinetics and -dynamics) and Class I MHC mediated antigen processing & presentation pathways to be predictive of infectious events.

Conclusion. Our data indicate that host genomic profile prediction of the risk of infectious events during induction therapy could be helpful in developing individualised supportive care and leukemic treatment strategies.

1. Introduction

Infections remain a significant challenge in treatment of childhood acute lymphoblastic leukemia (ALL) [24, 19, 2, 29, 4, 31] with known risk factors such as use of central venous catheters [3], mucositis [9], neutropenia [6] and treatment intensity [33]. Inherited genetic variation may influence the immune-inflammatory response in patients with compromised immune function such as children undergoing ALL treatment. A few studies testing single SNPs (single nucleotide polymorphisms) have shown associations between genetic variation and severity of infections during childhood leukemia treatment both within immunogenetics [17, 28, 20] and pharmacogenetics [10]. However, in children who were otherwise healthy prior to the diagnosis of ALL, severe infections are unlikely to be determined by single SNPs, since the complex factors precipitating these infections involve the interplay of multiple genes and their signalling molecules [15, 25]. In addition, significant single SNP associations generally have only mild effects on outcome [21]. Genome-wide association studies (GWAS) provide the opportunity to link new genes to well-defined outcomes such as response to treatment, toxicities, or other events. The biggest limitation of GWAS is that the majority of SNPs included on commercially available microarray platforms are located outside protein-coding genes or functionally annotated regions. Consequently, many significant SNPs in such studies are difficult to map to a specific gene, and causal relationships of significantly associated SNPs is often unclear [37, 11]. A hypothesis-driven extended candidate gene

[*]Corresponding author: Kjeld Schmiegelow, Department of Pediatrics, The Juliane Marie Centre, The University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark. Telephone +45 35451357, Fax +45 35454524, e-mail: kschmiegelow@rh.dk

approach including pathway-analysis and protein-protein interactions can provide more insight into the complex interplay between genes in complex diseases. We have also recently developed a cost-effective next-generation sequencing capture assay for SNP analysis allowing pooling and simultaneously genotyping of several samples for a large number of variants [36]. Aiming at exploring host genomic patterns of possible influence on infections during induction therapy, we expanded our earlier study on mannose binding lectin gene (MBL) polymorphisms and infections [19] to include approximately 34 000 target SNPs in candidate genes of possible relevance for childhood ALL treatment efficacy and toxicity.

2. Materials and Methods

The patient population has previously been described in detail [19]. Briefly, the included patients were i) diagnosed and treated according to the NOPHO ALL92 protocol [13, 32] for non-B ALL at University Hospital Rigshospitalet, Copenhagen, Denmark, in the period January 1, 1992, to December 31, 2000, and ii) between 1.0 -14.9 years of age at diagnosis. A total of 137 patients fulfilled these criteria and were included in the earlier study of MBL SNPs in relation to infections during the first 50 days of antileukemic therapy [19]. Of these, blood samples for extended SNP profiling were available for 72 patients. We excluded three patients because of too little amount of DNA resulting in 69 patients, who then constituted the study cohort (Table1). Patients were treated according to the following risk criteria: a) low risk criteria included WBC at diagnosis < 50x109/L and no other high-risk criteria; b) high risk criteria included at least one of the following: i) WBC > 50 x109/L, ii) T-cell disease, iii) a mediastinal mass, iv) t(9;22) and/or t(4;11), v) presence of CNS-disease and/or testicular involvement and/or lymphomatous leukemia, and vi) a M3 bone marrow day 15 and/or M2 bone marrow day 29.

Patients	Original cohort (n=137) (%)	Multiple SNP-cohort (n=69) (%) [a]	p-value for relation to "infectious event" [b] (48/69)	p-value for relation to "positive culture" [b] (23/69)
Sex			0.78	0.56
Female	50 (36.5)	18 (26.1)		
Male	87 (63.5)	51 (73.9)		
Age			<0.01	0.24
1-5	86 (62.8)	49 (71.0)		
6-10	32 (23.4)	14 (20.3)		
11-14	19 (13.9)	6 (8.7)		
Risk group			0.12	0.31
Low risk	95 (69.3)	53 (76.8)		
High risk	42 (30.7)	16 (23.2)		
Immunophenotype			0.71	0.47
Non-B cell	113 (82.5)	60 (87.0)		
T-cell	24 (17.5)	9 (13.0)		
Infectious event[c]				
Yes	99 (72.3)	48 (69.6)		
No	38 (27.7)	21 (30.4)		
Positive culture[c]				
Yes	43 (31.4)	23 (33.3)		
No	94 (68.6)	46 (66.7)		

Table 1: *Patient characteristics. a. Original SNP-cohort was 72 patients, but 3 patients were excluded due to too low DNA concentration. b. Association between patient characteristics and event (infectious event or positive culture), Pearson Chi-Square test or Fishers exact test, for the multiple SNP-cohort. c. At least 1 infectious event/positive blood culture during induction treatment.*

The four-weeks induction therapy and three-weeks post-induction period included in this study encompassed oral Prednisone 60 mg/m2/d days 1-36 and then tapered, weekly i.v. Vincristine 2.0 mg/m2 (six doses, maximum 2.0 mg per dose), i.v. Doxorubicin 40mg/m2 (days 1, 22 and 36; with the addition of one dose on day 8 for the high risk patients), Erwinia L-Asparaginase 30 000IE/m2 daily days 36-45, and age adjusted intrathecal Methotrexate (four doses) [13].

Fever was defined as a single temperature above $38.5\,^{\circ}\mathrm{C}$. Neutropenia was defined as an absolute neutrophil count (ANC) less than 0.5x109 /L. In this retrospective analysis, an infectious event was defined as the combination of fever or other signs of infection and initiation of antimicrobial therapy after the initiation of antileukemic therapy. In addition, a fraction of patients with an infectious event had positive blood cultures. Data for each new infectious event were collected for the first 50 days of treatment.

2.1 Clinical endpoints and grouping of patients

Clinical parameters were collected from the patient files and included all blood counts, days of neutropenia, febrile episodes, infectious events and microbiological data. For SNP association analyses patients were grouped twice. The first grouping included as cases the patients who had at least one infectious event (n=48) during the seven weeks of antileukemic therapy while the remaining 21 patients served as controls. The second grouping included cases from the first grouping, who in addition had at least one positive blood culture (n=23) with the remaining 46 patients serving as controls.

2.2 Candidate genes and SNP selection

Polymorphisms included in our multiple SNP analysis platform belonged to 2,350 known genes with possible relevance for childhood leukemia treatment efficacy and toxicity [36] covering the following areas/domains: i) pharmacogenetics, ii) immunogenetics, iii) apoptosis, iv) neurobiology, v) toxicity, vi) thrombosis, vii) cell cycle control genes and viii) DNA repair and mitosis. Targeted SNPs within candidate genes were selected based on their consequence on their transcript (according to Ensembl [14]

annotations): i) non-synonymous coding, ii) frame-shift coding, iii) regulatory region, iv) stop lost, v) stop gained, vi) splice site, vii) within non-coding gene, and viii) within mature micro-RNA. This resulted in selection of approximately 34 000 targeted SNPs assayed in this study.

2.3 Library preparation

Blood samples were obtained during morphological remission and DNA was extracted and purified by sodium chloride and ethanol precipitation. Library preparation was performed according to the SureSelect Target Enrichment System protocol version 1.2 April 2009 (Agilent Technologies, Santa Clara, CA, USA) with minor modifications as described earlier [36]. Briefly, three g genomic DNA was sheared by Covaris S2 System (Covaris Inc., Woburn, MA, USA) followed by DNA purification and end-repair. For multiplexing, patient specific barcodes of four bases were ligated to the DNA fragments. The library samples were then pooled in groups of three to eight samples and hybridized to custom designed baits (Sure Select Oligo Capture Libray, Agilent Technology) with two baits (or four in case of high priority SNPs) targeting each SNP. After PCR amplification the library were sequenced using Illumina HiSeq 2000 (Illumina Int., San Diego, CA, USA).

2.4 Sequence analysis

The high quality reads obtained from sequencing were mapped to the reference human genome build 37 (GRCh37) using Burrow-Wheelers Alignment (BWA) algorithm [22]. The alignment was refined by means of quality score recalibration and around indel realignment using Genome Analysis ToolKit package [26]. SNP calling was performed with SAMtools package [23] using default settings. The variants were filtered using the vcfutils.pl script from SAMtools package and only genotypes at a minimum of 10x sequencing depth were included in the analysis. SNP calling with our multiplexing technique have earlier been validated with PCR [36]. Variant annotation was done with Ensembl Variant Effect Predictor script [27].The data was converted into .ped and .map file formats readable by Plink [30] using vcftools package [8].

2.5 Single SNP association

The single SNPs associations to a risk of infection and chances of having a positive culture were performed by Fishers exact test implemented in PLINK. Only the SNPs with observed minor allele frequency (MAF) above 5% and at least 50% of non-missing genotypes above 10x sequencing depth were used in the analysis. The obtained p-values were adjusted for multiple testing with up to one million adaptive permutations [5]. The adjusted p-values were plotted on quantile-quantile (QQ) plots (Supplementary figure 1) using publicly available R script [1] and on Manhattan plot (Figure 1) using Circos software version 0.52 [18].

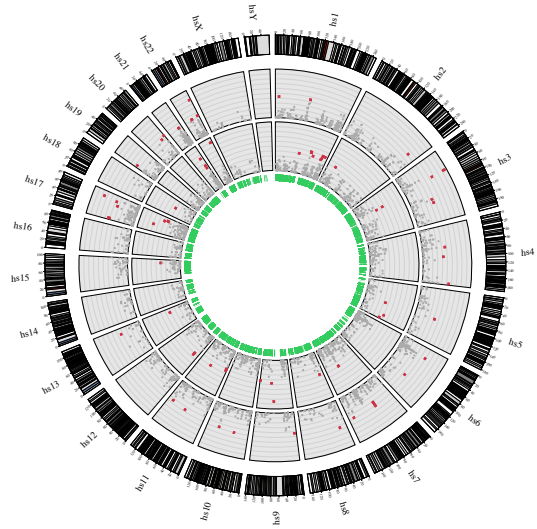


Figure 1: *Circular Manhattan plot showing the association of SNPs with 1) having an infectious event (inner gray-shaded ring), and 2) having a positive culture (outer gray-shaded ring). Each section (HS1, HS2, ...) represents the chromosomes (chr1, chr2,...). The radius (y-axis) represents the log10 of p-values and each SNP is plotted based on its position in the chromosome and as a function of its p-value. Red dots represents SNPs with p-values < 0.0001. Green tickmarks represent positions of baits used for target capture.*

2.6 Pathway analysis

All functional (non-synonymous coding, frame-shift coding, stop codon and splice site) SNPs genotyped in this study with MAF > 0.01 residing in the pathways genes were retrieved for Reactome pathways [16] excluding the top two pathway levels. In order to discover pathways

related to outcome, machine learning was performed to predict outcome using combinations of SNPs within pathways. Number of SNPs per pathway ranged from 1 to 82 SNPs, and each SNP was encoded by three values between 0 and 1 corresponding to likelihood of each genotype calculated from VCF file produced by SAMtools [23]. Missing variants were encoded as observed population frequencies for the three genotypes. Associations to infectious event and positive blood culture were performed by training artificial neural networks on subsets of SNPs from each pathway with 3-fold cross validation. For each pathway all combinations of up to three SNPs were assessed, and the tested combinations of SNPs were ranked by Matthews correlation coefficient (MCC) calculated as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives. For each resulting best combination of SNPs for each pathway the neural network parameters were optimized by testing multiple settings of hidden neurons and training cycles. Pathways were then ranked by MCC of the best combination of SNPs for each pathway. For all best combinations the area under receiver-operator curve (AUC) was calculated with 95% confidence intervals with the 'pROC' R package.

2.7 CART

Classification and regression tree (CART) analyses were performed using rpart R package applying 3-fold cross-validation [35] using the genotypes of the SNPs associated to risk of infection and positive culture together with the patients' age, sex, risk group, immunophenotype and white blood cell count (WBC) at diagnosis. Only the SNPs with permutation corrected P-values below 0.01 were used in the analysis. The same analysis was repeated including also pathway-based predictions achieved from the top ten artificial neural network classifiers from pathway analysis.

2.8 Additional statistics

In addition to the above mentioned bioinformatics approaches, Chi-Square or Fishers Exact Test were used for univariable analysis. For "time-to-event" analysis, Kaplan Meyer analysis was done for univariable comparison with log-rank test and proportional Hazard ratios calculated for multiple regression analysis. The software package used was IBM SPSS statistics version 20.

2.9 Ethics

The study was approved by the Committee for research ethics in the Danish Region H and performed in accordance of Declaration of Helsinki.

3. Results

Of the 69 included patients in the SNP cohort, 48 (70%) experienced at least one infectious event during the 7 weeks induction period (Table I). Of these, 23 (33A total of 103 and 94 SNPs were found to be associated with infectious event and positive culture, respectively ($p<0.05$ for all associations) (Supplementary tables I and II). The QQ plots for the single SNP association analysis followed the null distribution, but showed a few SNPs above the expected distributions, suggesting true biological signals (Supplementary figure 1). Manhattan plots for both analyses indicate that several loci were associated to both risk of having an infection and positive culture (Figure 1). The most significantly associated SNP was a synonymous coding SNP rs11033797 in *ORF51F1*, while the top SNP associated to positive blood culture was a non-synonymous coding SNP rs12632456 in *FLNB* gene. Pathway analysis identified 'GPCR downstream signaling' (MCC = 0.72, AUC = 0.88), 'Bile acid and bile salt metabolism' (MCC = 0.71, AUC = 0.89), and 'Class I MHC mediated antigen processing & presentation' (MCC = 0.68, AUC = 0.83) among the pathways most predictive of an infectious event. 'Interferon Signaling' (MCC = 0.70, AUC = 0.85), 'Rho GTPase cycle' (MCC = 0.62, AUC = 0.86) and 'G alpha (i) signaling events' (MCC = 0.60, AUC = 0.85) pathways were the most predictive of having a positive blood culture. The top ten pathways together with selected combinations of SNPs for risk of infectious event are listed in Supplementary Table III and for positive cultures in Supplementary Table IV.

CART analysis demonstrated rs11033797 (SNP1, *OR51F1*), rs2835265 (SNP2, *CBR1*), rs28627172 (SNP3, *POLDIP3*) and rs1129844/CM072923 (SNP4, *CCL11*) to be highly predictive of infectious event (Figure 2). Patients were classified as "SNP profile-positive" (P+) according to the following SNP-pattern: either i) SNP1(GA, GG) or ii) SNP1(AA) and SNP2(TC), or iii) SNP1(AA) and SNP2(CC) and SNP3(GG) and SNP4(GG). The remaining patients were classified as "SNP profile negative" (P-). Data for all four SNPs in the profile were available for 61 out of 69 patients. Of these 61 patients, 45 experienced at least one infectious event. Of the P+ patients, 40 (97.6%) out of 41 experienced an event compared to only 5 (25%) out of the 20 P- patients (Figure 3, Table III). In a time-to-event analysis, the risk of having an infectious event for P+ patients compared to P- patients was 8.9 (95% CI:

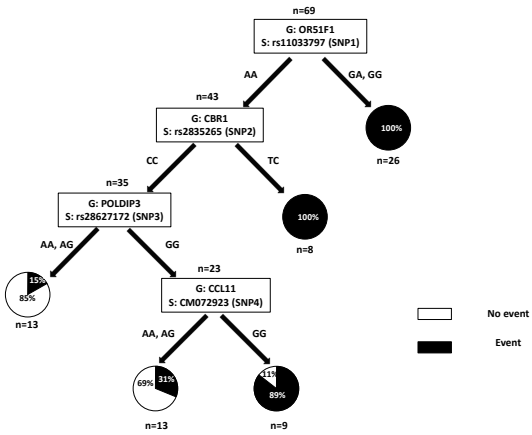


Figure 2: *CART (Classification and regression tree) diagram showing the most predictive SNP-sequences (all SNPs p<0.01). Text-boxes: G: gene S: SNP (SNPs are termed SNP1-4 for simplicity).*

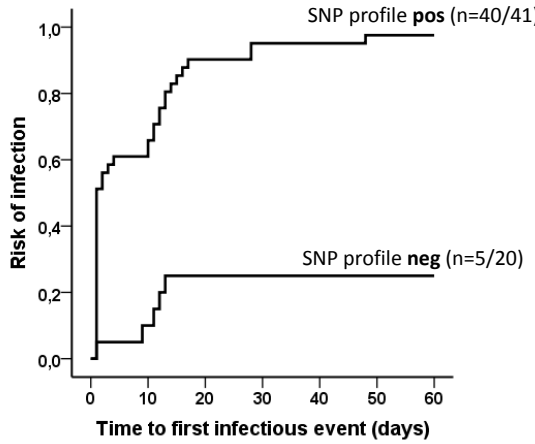


Figure 3: *SNP profile and time to first infectious event. Total number of patients: 61, total number of events: 45. The figure illustrates the risk of having an infectious event with positive (97.6% + 2.4%, 40/41) vs. negative (25% + 9.7%, 5/20) SNP profile (p<0.001,Log Rank).*

3.2-23.5, Cox regression). Adjusting for the mean absolute neutrophil counts and age did not change the risk significantly. The age group 1-5 years came out significantly in simple regression with a risk of 9.4 (95% CI: 1.3-69.0), but after adjustment for SNP-profile and ANC, age lost statistical significance (Table III). For the "culture positive" patients, CART analysis was less predictive and SNPs rs12632456 (*FLNB*) and rs1171218 (*TOPBP1*) were predictive for only 10 of 23 (43%) culture positive patients with 77% accuracy (Figure 2).

CART analysis including pathway profiles demonstrated a combination of variations in 'GPCR downstream signaling', 'Bile acid and salt metabolism' and 'Class I MHC antigen processing & presentation' to be highly predictive of infectious event (Supplementary Figure 1), while combinations of variations within 'Interferon signaling' and 'Platelet Aggregation (Plug Formation)' pathways were predictive of a positive blood culture (Supplementary Figure 2).

Risk factor	Events n=45	All patients n=61	HR (95% CI) Simple regr.	Adjusted HR (95% CI) Multiple regr.
SNP profile				
Positive	40	41	**8.9 (3.2-23.5)***	**7.5 (2.8-20.1)***
Negative	5	20	1.0	1.0
wmANC				
<median	25	32	1.2 (0.7-2.2)	0.8 (0.5-1.5)
>median	20	29	1.0	1.0
Age				
1-5	38	43	**9.4 (1.3-69.0)***	4.9 (0.6-38.1)
6-10	6	13	2.7 (0.3-22.7)	1.6 (0.2-13.3)
11-14	1	5	1.0	1.0

HR: Hazard ratio
wmANC: weighted mean of ANC
Only patients with sufficient SNP profile sequenced was included
*significant

Table 2: *Multiple regression analysis showing risk factors for having an infectious event.*

4. Discussion

The cure rates of childhood ALL are approaching more than 85-90%, but many of the 10-15% of patients that die, do so from treatment related toxicity rather than from active cancer. Since relapse rates are decreasing because of better risk adapted but also more intensive therapy, focus on toxic events has become increasingly essential. Among patients dying from toxicity, infections are the most common cause, and the majority of these occur during the first months of treatment [24, 29]. It remains unknown, why some patients develop severe, even life-threatening infections during treatment, while others, within the same sex, age and risk group only experience mild infections. Many factors influence the risk of severe infections during ALL treatment (Figure 4), however, the role of inherited genetic factors has to date not been investigated extensively.

The present study strongly indicates that common host genomic variants that influence immune function and drug disposition and effector mechanisms may play a critical role. However, infectious events constitute a clinically heterogeneous group spanning from mild clinical episodes of fever of unknown origin to severe infectious episodes requiring intensive care. Accordingly, the retrospective nature of this study and the definition of an infectious event (fever or other signs of infection and start of antibi-

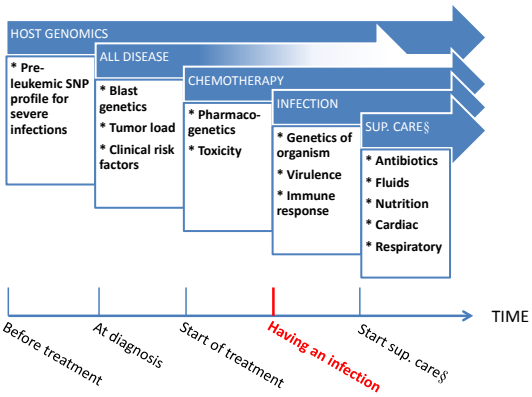


Figure 4: *Factors influencing infectious events during ALL treatment. §= Supportive care*

otics) may have both under- and overestimated the impact of host genomics.

Because of the complex nature of infections, a multiple gene and SNP profiling approach is more likely to yield biologically relevant results when compared to single SNP analysis. Importantly, our findings of 103 and 94 SNPs significantly associated with the risk of an infectious event or a positive culture, emphasize that application of such a candidate gene approach can yield potentially relevant results, not least when the SNP data are integrated into pathway explorations. Still, the results of this study need to be validated in independent cohorts.

Integrative analysis of effects mediated by multiple SNPs grouped by their function, e.g. acting in the same biological pathway, can provide robust results identifying the most important underlying biological mechanisms. The two pathways most predictive of infectious event 'GPCR downstream signaling' and 'Bile acid and salt metabolism' are connected to steroid drugs that constitute an important component of induction therapy [12]. Variants in the second pathway are within genes involved in steroid metabolism, while GPCR signalling reflects the mode of action of those drugs [34]. Finally, 'Class I MHC antigen processing & presentation' is a critical pathway in response to viral and other infections, as are the two pathways most predictive of positive blood culture 'Interferon signaling' and 'Rho GTPase cycle'.

Although the hierarchical approach of CART analysis reduces the true complexity of the biological and environmental (including instrumental) factors leading to infections, its simplicity makes the risk profile data more comprehensible and clinically applicable than complex pathway interactions. The CART approach identified a SNP risk profile, which predicted the occurrence of an infectious event for 40 (89%) out of 45 patients with 98% accuracy. The SNPs selected by CART algorithm to collectively predict the infection risk resided in genes involved in olfactory functions (*OR51F1*), steroid and anthracycline metabolism (*CBR1*), cell growth (*POLDIP3*) and chemokine signalling (*CCL11*). When including pathway profiles in the analysis, variants affecting GPCR downstream signalling (including another olfactory gene *OR51T1*) and steroid metabolism ('Bile acid and salt metabolism') were most predictive of outcome. In case of positive blood culture, CART analysis identified the SNP in the *FLNB* gene to best separate cases from controls, which was also the top associated SNP in Fishers exact test. CART analysis including pathway profiles indicated 'Interferon signalling' (including the variant in *FLNB* gene) and 'Platelet Aggregation (Plug Formation)' to predict the phenotype best.

The results of this study reflect the suspected pharmacogenomic and immune system mechanisms of infections. Glucocorticosteroids act by suppressing the immune system, and the risk of infection has been shown to increase with dose and duration of treatment with those drugs [7]. Class I MHC antigen presentation mediates recognition of infectious pathogens, and this process is upregulated by interferon signalling which triggers the immune system response, while aggregation of platelets is a mechanism activated in response to inflammatory reactions caused by e.g. infection. These results suggest that patients with increased risk of infectious events are likely to experience higher systemic exposure of glucocorticosteroids and/or have impaired certain immune system functions due to their genomic background.

Our study had several limitations, including a small number of patients, and a risk of false positive results. Additionally, investigations of a set of candidate polymorphisms, even when conducted on a large scale, a priori exclude identification of other important biological mechanisms. Such hypothesis generating exploration will require less supervised genome wide analyses such as GWAS and exome or whole genome sequencing. Finally, DNA was available for only 60% of the originally patient cohort and the influence of this bias remains unclear.

Despite the small size of the cohort, the presented results indicate that host genomic profile can predict the risk of infectious events during induction therapy in children with ALL. Such knowledge might be helpful in adapting more individualised supportive care and treatment protocols. Furthermore, the study design used here can be used for many other clinical endpoints, for instance other toxicities such as toxic deaths, pancreatitis, thrombosis, and osteonecrosis. Such candidate SNP profiling and association studies are ongoing as part of the Nordic/Baltic ALL collaboration on childhood and adult ALL, and results could be applied for future altered treatment strategies by i) reducing treatment intensity, ii) use of antimicrobial prophylaxis, and iii) administration of immune therapy (immunoglobulins or specific immune therapy).

Conflict of interest

The authors declare no conflict of interest.

Author Contribution

Bendik Lund and Kjeld Schmiegelow planned the study. Ramneek Gupta, Agata Wesolowska-Andersen, Louise Borst and Kjeld Schmiegelow identified the relevant genes and designed the SNP profiling approach. Bendik Lund and Agata Wesolowska-Andersen wrote the manuscript, performed data analysis, and interpreted the data. Birgitte Lausen was responsible for collection of the clinical data. Louise Borst and Kirsten Kørup Rasmussen performed the laboratory work. Helge Klungland and Ramneek Gupta supervised data analysis. All authors provided critical input to the project and manuscript and approved the final manuscript.

Acknowledgements

This study received financial support from the Danish Childhood Cancer Foundation; the Danish Medical Research Council for Health and Disease; The Danish Cancer Society; The Novo Nordic Foundation; the Liaison committee Central Norway Regional Health Authority, Norwegian University of Science and Technology (NTNU); Støtteforeningen for kreftsyke barn, Trøndelag.

References

[1] "Getting Genetics Done - http://gettinggeneticsdone.blogspot.com/".

[2] S. Afzal, M. Ethier, L. Dupuis, L. Tang, A. Punnett, S. Richardson, U. Allen, O. Abla, and L. Sung, "Risk factors for infection-related outcomes during induction therapy for childhood acute lymphoblastic leukemia", *The Pediatric infectious disease journal*, Vol. 28, No. 12, p. 1064, 2009.

[3] R. Allen, M. Holdsworth, C. Johnson, C. Chavez, R. Heideman, G. Overturf, D. Lemon, W. Hunt, and S. Winter, "Risk determinants for catheter-associated blood stream infections in children and young adults with cancer", *Pediatric blood & cancer*, Vol. 51, No. 1, pp. 53–58, 2008.

[4] L. Bailey, A. Reilly, and S. Rheingold, "Infections in pediatric patients with hematologic malignancies", In *Seminars in hematology*, Vol. 46, pp. 313–324, Elsevier, 2009.

[5] J. Besag and P. Clifford, "Sequential monte carlo p-values", *Biometrika*, Vol. 78, No. 2, pp. 301–304, 1991.

[6] E. Castagnola, V. Fontana, I. Caviglia, S. Caruso, M. Faraci, F. Fioredda, M. Garrè, C. Moroni, M. Conte, G. Losurdo, et al., "A prospective study on the epidemiology of febrile episodes during chemotherapy-induced neutropenia in children with cancer or after hemopoietic stem cell transplantation", *Clinical infectious diseases*, Vol. 45, No. 10, pp. 1296–1304, 2007.

[7] M. Cutolo, B. Seriolo, C. Pizzorni, M. Secchi, S. Soldano, S. Paolino, P. Montagna, and A. Sulli, "Use of glucocorticoids and risk of infections", *Autoimmunity reviews*, Vol. 8, No. 2, pp. 153–155, 2008.

[8] P. Danecek, A. Auton, G. Abecasis, C. Albers, E. Banks, M. DePristo, R. Handsaker, G. Lunter, G. Marth, S. Sherry, et al., "The variant call format and VCFtools", *Bioinformatics*, Vol. 27, No. 15, pp. 2156–2158, 2011.

[9] L. Elting, C. Cooksley, M. Chambers, S. Cantor, E. Manzullo, and E. Rubenstein, "The burdens of cancer therapy", *Cancer*, Vol. 98, No. 7, pp. 1531–1539, 2003.

[10] D. Erdélyi, E. Kámory, A. Zalka, Á. Semsei, B. Csókay, H. Andrikovics, A. Tordai, G. Borgulya, E. Magyarosy, I. Galántai, et al., "The role of ABC-transporter gene polymorphisms in chemotherapy induced immunosuppression, a retrospective study in childhood acute lymphoblastic leukaemia", *Cellular immunology*, Vol. 244, No. 2, pp. 121–124, 2006.

[11] E. Feingold, P. Good, M. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F. Collins, T. Gingeras, D. Kampa, E. Sekinger, et al., "The ENCODE (ENCyclopedia of DNA elements) project", *Science*, Vol. 306, No. 5696, pp. 636–640, 2004.

[12] P. Gaynon and A. Carrel, "Glucocorticosteroid therapy in childhood acute lymphoblastic leukemia.", *Advances in experimental medicine and biology*, Vol. 457, p. 593, 1999.

[13] G. Gustafsson, K. Schmiegelow, E. Forestier, N. Clausen, A. Glomstein, G. Jonmundsson, L. Mellander, A. Mäkipernaa, R. Nygaard, U. Saarinen-Pihkala, et al., "Improving outcome through two decades in childhood ALL in the Nordic countries: the impact of high-dose methotrexate in the reduction of CNS irradiation. Nordic Society of Pediatric Haematology and Oncology (NOPHO).", *Leukemia: official journal of the Leukemia Society of America, Leukemia Research Fund, UK*, Vol. 14, No. 12, p. 2267, 2000.

[14] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, et al., "The Ensembl genome database project", *Nucleic acids research*, Vol. 30, No. 1, pp. 38–41, 2002.

[15] R. Huttunen and J. Aittoniemi, "New concepts in the pathogenesis, diagnosis and treatment of bacteremia and sepsis", *Journal of Infection*, Vol. 63, No. 6, pp. 407–419, 2011.

[16] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, et al., "Reactome: a knowledgebase of biological pathways", *Nucleic acids research*, Vol. 33, No. suppl 1, pp. D428–D432, 2005.

[17] E. Kidas, A. Möricke, R. Beier, K. Welte, M. Schrappe, M. Stanulla, and L. Grigull, "Genetic polymorphisms of the lymphotoxin alpha gene are associated with increased risk for lethal infections during induction therapy

for childhood acute leukemia: a case-control study", *International journal of hematology*, Vol. 89, No. 5, pp. 584–591, 2009.

[18] M. Krzywinski, J. Schein, İ. Birol, J. Connors, R. Gascoyne, D. Horsman, S. Jones, and M. Marra, "Circos: an information aesthetic for comparative genomics", *Genome research*, Vol. 19, No. 9, pp. 1639–1645, 2009.

[19] B. Lausen, K. Schmiegelow, B. Andreassen, H. Madsen, and P. Garred, "Infections during induction therapy of childhood acute lymphoblastic leukemia–no association to mannose-binding lectin deficiency", *European journal of haematology*, Vol. 76, No. 6, pp. 481–487, 2006.

[20] T. Lehrnbecher, T. Bernig, M. Hanisch, U. Koehl, M. Behl, D. Reinhardt, U. Creutzig, T. Klingebiel, S. Chanock, and D. Schwabe, "Common genetic variants in the interleukin-6 and chitotriosidase genes are associated with the risk for serious infection in children undergoing therapy for acute myeloid leukemia", *Leukemia*, Vol. 19, No. 10, pp. 1745–1750, 2005.

[21] T. Lesnick, S. Papapetropoulos, D. Mash, J. Ffrench-Mullen, L. Shehadeh, M. De Andrade, J. Henley, W. Rocca, J. Ahlskog, and D. Maraganore, "A genomic pathway approach to a complex disease: axon guidance and Parkinson disease", *PLoS genetics*, Vol. 3, No. 6, p. e98, 2007.

[22] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows–Wheeler transform", *Bioinformatics*, Vol. 26, No. 5, pp. 589–595, 2010.

[23] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al., "The sequence alignment/map format and SAMtools", *Bioinformatics*, Vol. 25, No. 16, pp. 2078–2079, 2009.

[24] B. Lund, A. Åsberg, M. Heyman, J. Kanerva, A. Harila-Saari, H. Hasle, S. Söderhäll, Ó. Jónsson, S. Lydersen, and K. Schmiegelow, "Risk factors for treatment related mortality in childhood acute lymphoblastic leukaemia", *Pediatric blood & cancer*, Vol. 56, No. 4, pp. 551–559, 2011.

[25] J. Marshall, J. Vincent, M. Fink, D. Cook, G. Rubenfeld, D. Foster, C. Fisher Jr, E. Faist, and K. Reinhart, "Measures, markers, and mediators: toward a staging system for clinical sepsis. A report of the Fifth Toronto Sepsis Roundtable, Toronto, Ontario, Canada, October 25-26, 2000", *Critical care medicine*, Vol. 31, No. 5, p. 1560, 2003.

[26] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, et al., "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data", *Genome research*, Vol. 20, No. 9, pp. 1297–1303, 2010.

[27] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor", *Bioinformatics*, Vol. 26, No. 16, pp. 2069–2070, 2010.

[28] K. Miedema, E. Te Poele, W. Tissing, D. Postma, G. Koppelman, A. de Pagter, W. Kamps, B. Alizadeh, H. Boezen, and E. de Bont, "Association of polymorphisms in the TLR4 gene with the risk of developing neutropenia in children with leukemia", *Leukemia*, Vol. 25, No. 6, pp. 995–1000, 2011.

[29] C. Prucker, A. Attarbaschi, C. Peters, M. Dworzak, U. Pötschger, C. Urban, F. Fink, B. Meister, K. Schmitt, O. Haas, et al., "Induction death and treatment-related mortality in first remission of children with acute lymphoblastic leukemia: a population-based analysis of the Austrian Berlin-Frankfurt-Münster study group", *Leukemia*, Vol. 23, No. 7, pp. 1264–1269, 2009.

[30] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. De Bakker, M. Daly, et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses", *The American Journal of Human Genetics*, Vol. 81, No. 3, pp. 559–575, 2007.

[31] J. Rahiala, M. Perkkiö, and P. Riikonen, "Infections occurring during the courses of anticancer chemotherapy in children with ALL: a retrospective analysis of 59 patients", *Pediatric Hematology-Oncology*, Vol. 15, No. 2, pp. 165–174, 1998.

[32] K. Schmiegelow, E. Forestier, M. Hellebostad, M. Heyman, J. Kristinsson, S. Söderhäll, and M. Taskinen, "Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia", *Leukemia*, Vol. 24, No. 2, pp. 345–354, 2009.

[33] L. Sung, A. Gamis, T. Alonzo, A. Buxton, K. Britton, J. DeSwarte-Wallace, and W. Woods, "Infections and association with different intensity of chemotherapy in children with acute myeloid leukemia", *Cancer*, Vol. 115, No. 5, pp. 1100–1108, 2009.

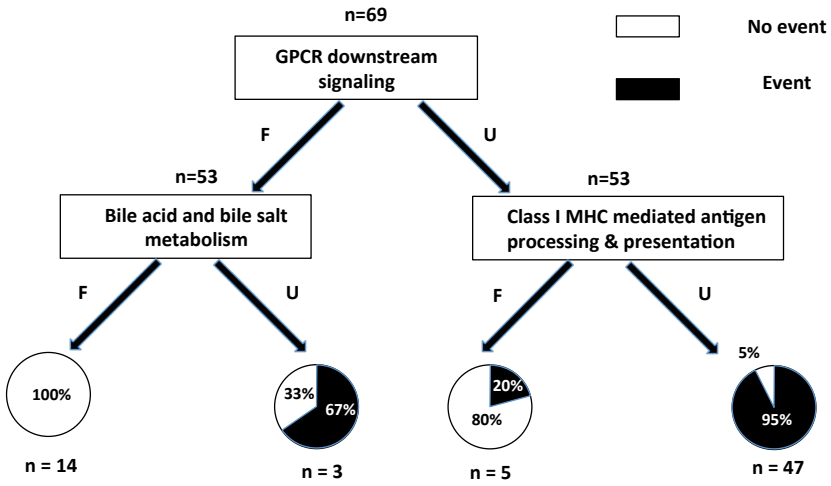
[34] J. Tasker, S. Di, and R. Malcher-Lopes, "Rapid glucocorticoid signaling via membrane-associated receptors", *Endocrinology*, Vol. 147, No. 12, pp. 5549–5556, 2006.

[35] T. Therneau and E. Atkinson, An introduction to recursive partitioning using the RPART routines, Technical report, Technical Report 61, Section of Biostatistics, Mayo Clinic, Rochester. URL http://www. mayo. edu/hsr/techrpt/61. pdf, 1997.

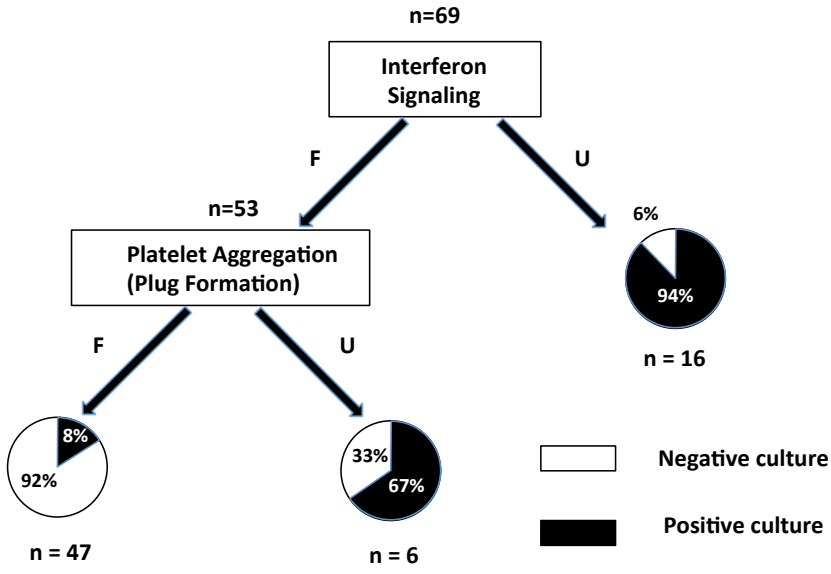
[36] A. Wesolowska, M. Dalgaard, L. Borst, L. Gautier, M. Bak, N. Weinhold, B. Nielsen, L. Helt, K. Audouze, J. Nersting, et al., "Cost-effective multiplexing before capture allows screening of 25 000 clinically relevant SNPs in childhood acute lymphoblastic leukemia", *Leukemia*, Vol. 25, No. 6, pp. 1001–1006, 2011.

[37] J. Yang, C. Cheng, W. Yang, D. Pei, X. Cao, Y. Fan, S. Pounds, G. Neale, L. Treviño, D. French, et al., "Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia", *JAMA: the journal of the American Medical Association*, Vol. 301, No. 4, pp. 393–403, 2009.

Supplementary Figure 1. CART analysis for risk of infectious events including pathway profiles. F = favorable profile, U = unfavorable profile.



Supplementary Figure 2. CART analysis for risk of positive blood culture including pathway profiles. F = favorable profile, U = unfavorable profile.



Supplementary Table 1. SNPs associated with risk of infectious event with permutation corrected *P*-values < 0.01

rsID	Perm p-val	Genotypes – no event	Genotypes - event	SNP consequence	HGNC
rs11033797	6.00E-06	GG:0;GA:0;AA:20;00:1	GG:1;GA:25;AA:22;00:0	SYNONYMOUS_CODING	*OR51F1*
rs28627172	0.0001	AA:2;AG:8;GG:8;00:3	AA:0;AG:5;GG:40;00:3	NON_SYNONYMOUS_CODING	*POLDIP3*
rs35158358	0.0002	-/-:0;-/+C:9;+C/+C:10;00:2	+C/+C:6;+C/-:26;-/-:12/00:4	WITHIN_NON_CODING_GENE	*GDA*
rs73197348	0.0007	CC:0;CT:7;TT:13;00:1	CC:0;CT:2;TT:46;00:0	REGULATORY_REGION	*RUNX1*
rs1129844, CM072923	0.001	AA:3;AG:7;GG:6;00:5	AA:0;AG:10;GG:34;00:4	NON_SYNONYMOUS_CODING	*CCL11*
rs2835265	0.004	TT:0;TC:0;CC:20;00:1	TT:0;TC:14;CC:34;00:0	NON_SYNONYMOUS_CODING	*CBR1*
rs11409972	0.004	+T/+T:1;+T/-:7; -/-:10;00:3	+T/+T:9;+T/-:24;-/-:12;00:3	REGULATORY_REGION	*MAGI1*
rs67155431	0.005	-/-:1;-/+A:4;+A/+A:7;00:9	-/-:0;-/+A:4;+A/+A:38/00:6	SPLICE_SITE	*SLCO1A2*
rs10653020	0.007	+CATC/+CATC:0;+CATC/-:1;-/-:17;00:3	+CATC/+CATC:0;+CATC/-:17;-/-:27;00:4	WITHIN_NON_CODING_GENE	*ACE*
rs1130435	0.007	TT:1;TC:5;CC:10;00:5	TT:8;TC:24;CC:10;00:6	NON_SYNONYMOUS_CODING	*FABP6*
rs76959009	0.007	CC:0;CT:2;TT:18;00:1	CC:0;CT:20;TT:26;00:2	WITHIN_NON_CODING_GENE	*C1orf112*
rs1136410, CM042761	0.007	GG:0;GA:8;AA:3;00:10	GG:0;GA:10;AA:27;00:11	NON_SYNONYMOUS_CODING	*PARP1*
rs35710857	0.007	TT:0;TC:0;CC:21;00:0	TT:0;TC:13;CC:35;00:0	SYNONYMOUS_CODING	*CBR1*
rs2292572	0.007	TT:0;TG:2;GG:18;00:1	TT:5;TG:15;GG:28;00:0	REGULATORY_REGION	*GAB2*
rs2275287	0.007	TT:1;TC:4;CC:9;00:7	TT:8;TC:27;CC:9;00:4	SPLICE_SITE	*RYR2*
rs10841795	0.007	GG:0;GA:5;AA:10;00:6	GG:0;GA:3;AA:43;00:2	NON_SYNONYMOUS_CODING	*SLCO1A2*
rs12666401	0.007	AA:0;AG:0;GG:14;00:7	AA:0;AG:14;GG:27;00:7	REGULATORY_REGION	*SHFM1*
rs5367	0.008	GG:0;GA:2;AA:17;00:2	GG:0;GA:20;AA:25;00:3	SPLICE_SITE	*SELE*
rs12133666	0.008	TT:0;TA:2;AA:19;00:0	TT:0;TA:20;AA:28;00:0	WITHIN_NON_CODING_GENE	*SELE*
rs3917441	0.008	CC:0;CG:2;GG:18;00:1	CC:0;CG:20;GG:28;00:0	REGULATORY_REGION	*SELE*
rs77115118	0.008	TT:0;TA:2;AA:19;00:0	TT:0;TA:20;AA:28;00:0	WITHIN_NON_CODING_GENE	*SELE*
rs62471402	0.008	GG:0;GC:0;CC:15;00:6	GG:0;GC:14;CC:33;00:1	WITHIN_NON_CODING_GENE	*SHFM1*
rs4877837	0.009	AA:0;AG:6;GG:3;00:12	AA:0;AG:7;GG:26;00:15	REGULATORY_REGION	*SLC28A3*
rs2306825	0.009	AA:0;AG:1;GG:16;00:4	AA:0;AG:17;GG:28;00:3	SYNONYMOUS_CODING, SPLICE_SITE	*PSD3*

Supplementary Table 2. SNPs associated with positive blood culture with permutation corrected *P*-values < 0.01.

rsID	Perm p-val	Genotypes – positive culture	Genotypes – no positive culture	SNP consequence	HGNC
rs12632456	7.36E-05	AA:2;AG:14;GG:7;00:0	AA:0;AG:10;GG:35;00:1	NON_SYNONYMOUS_CODING	*FLNB*
rs12638356	9.12E-05	GG:3;GA:14;AA:6;00:0	GG:0;GA:12;AA:34;00:0	REGULATORY_REGION	*FLNB*
rs61748935	0.0004	AA:1;AG:5;GG:15;00:2	AA:0;AG:0;GG:42;00:4	NON_SYNONYMOUS_CODING	*C22orf40*
rs8464	0.0005	AA:1;AC:11;CC:11;00:0	AA:0;AC:5;CC:40;00:1	INTERGENIC, REGULATORY_REGION	*NA*
rs8640	0.001	TT:1;TC:14;CC:6;00:2	TT:0;TC:11;CC:28;00:7	SYNONYMOUS_CODING	*FLNB*
rs74207980, rs9270773	0.003	GG:1;GA:12;AA:7;00:3	GG:0;GA:10;AA:28;00:8	INTERGENIC	*NA*
rs2070687	0.004	GG:0;GC:4;CC:19;00:0	GG:1;GC:21;CC:21;00:3	SPLICE_SITE	*SFTPC*
rs35709976	0.004	-/-:0;-/+TA:2;+TA/+TA:21;00:0	-/-:0;-/+TA:19;+TA/+TA:26;00:1	REGULATORY_REGION	*F11*
rs6065, CM032257	0.005	TT:0;TC:11;CC:12;00:0	TT:0;TC:6;CC:37;00:3	NON_SYNONYMOUS_CODING	*GP1BA*
rs7577978	0.005	AA:2;AG:6;GG:8;00:7	AA:0;AG:4;GG:22;00:20	DOWNSTREAM	*ATIC*
rs238239	0.006	CC:3;CT:11;TT:9;00:0	TT:4;TC:27;CC:14;00:1	NON_SYNONYMOUS_CODING	*ENO3*
rs10949870	0.006	AA:3;AG:14;GG:4;00:2	AA:0;AG:16;GG:17;00:13	REGULATORY_REGION	*ZNF727*
rs59337853	0.007	-/-:2;-/T:4/TT:15;00:2	-/-:0;-/T:2;TT:39;00:5	REGULATORY_REGION	*ALOX5*
rs35187157	0.008	-/-:1;-/+TT:8;+TT/+TT:14;00:0	-/-:6;-/+TT:28;+TT/+TT:11;00:1	REGULATORY_REGION	*TRBV30*
rs3787537	0.008	TT:2;TC:11;CC:10;00:0	TT:0;TC:12;CC:32;00:2	REGULATORY_REGION	*SLCO4A1*
rs9332119	0.009	CC:0;CG:10;GG:11;00:2	CC:0;CG:6;GG:35;00:5	WITHIN_NON_CODING_GENE	*CYP2C9*
rs4298	0.009	TT:1;TC:5;CC:13;00:4	TT:0;TC:2;CC:31;00:13	SYNONYMOUS_CODING	*ACE*
rs6958588	0.009	TT:3;TC:13;CC:4;00:3	TT:0;TC:19;CC:17;00:10	WITHIN_NON_CODING_GENE	*ZNF727*
rs976002, rs139555919	0.009	GG:4;GA:10;AA:8;00:1	GG:1;GA:15;AA:27;00:3	NON_SYNONYMOUS_CODING	*TMPRSS11E*
rs2228539	0.009	CC:4;CT:11;TT:8;00:0	CC:2;CT:13;TT:31;00:0	NON_SYNONYMOUS_CODING	*EMR1*
rs11712186	0.0098	CC:1;CT:8;TT:6;00:8	CC:0;CT:7;TT:22;00:17	REGULATORY_REGION	*TOPBP1*

Supplementary Table 3. Top ten Reactome pathways predictive of infectious event.

MCC	AUC (CI%95)	Pathway name	SNPs / genes
0.72	0.88 (0.75 - 0.97)	GPCR downstream signaling	chr2:178528629 (PDE11A), rs4762,CM920009 (AGT), rs78644275 (OR51T1)
0.71	0.89 (0.81 – 0.98)	Bile acid and bile salt metabolism	CM014711 (HSD17B4), rs11045681 (SLCO1B7), rs41272687,CM005424 (CYP27A1)
0.70	0.91 (0.84 – 0.99)	GPCR ligand binding	rs970388 (GABBR2), rs17611 (C5), rs4762,CM920009 (AGT)
0.70	0.91 (0.84 – 0.99)	Peptide ligand-binding receptors	rs2277984 (C3), rs17611 (C5), rs4762,CM920009 (AGT)
		Class A/1 (Rhodopsin-like receptors)	
0.68	0.83 (0.72 – 0.94)	Class I MHC mediated antigen processing & presentation	rs11558955 (RAD23A), rs4036 (TCEB2), rs4673,CM983302 (CYBA)
0.68	0.79 (0.66 – 0.93)	G alpha (q) signalling events	CM012741 (CASR), rs145073237 (XCL1), rs4762,CM920009 (AGT)
0.66	0.89 (0.81 – 0.97)	Response to elevated platelet cytosolic Ca2+	rs216902 (VWF), rs2562830 (TTN), rs6023 (F5)
		Platelet degranulation	
0.62	0.77 (0.63 – 0.90)	G1/S Transition	rs2071467 (TAP2), rs45568137 (PPP2R1B), rs61732929 (POLE)
		Mitotic G1-G1/S phases	
		Mitotic M-M/G1 phases	
0.61	0.77 (0.63 – 0.90)	Gastrin-CREB signalling pathway via PKC and MAPK	CM012741 (CASR), rs148047905 (MMP3), rs4762,CM920009 (AGT)
0.55	0.80 (0.69 – 0.91)	Cytochrome P450 - arranged by substrate type	rs1126545 (CYP2C18), rs58871670 (CYP2B6), rs41272687,CM005424 (CYP27A1)
		Phase 1 - Functionalization of compounds	

Supplementary Table4. Top ten Reactome pathways predictive of positive/negative culture.

MCC	AUC (CI%95)	Pathway name	SNPs / genes
0.70	0.85 (0.73 – 0.96)	Interferon Signaling	rs1049069 (HLA-DQB1), rs12632456 (FLNB), rs146778723 (HLA-C)
0.62	0.86 (0.77 – 0.95)	Rho GTPase cycle	rs11800462 (TNFRSF25), rs1801284,CM982053 (HMHA1), rs2061821 (AKAP13)
0.60	0.85 (0.75 – 0.95)	G alpha (i) signalling events	rs114642578 (UBD), rs11575580 (IL11RA), rs61745073 (ADCY4)
0.58	0.84 (0.74 – 0.94)	Interleukin-2 signaling	rs11256369 (IL2RA), rs117805308 (CSF2RB), rs290223 (SYK)
		Interleukin-3, 5 and GM-CSF signaling	
		Signaling by Interleukins	
0.53	0.83 (0.72 – 0.93)	Glucose metabolism	rs238239 (ENO3), rs11208257 (PGM1), rs6065,CM032257 (GP1BA)
0.50	0.77 (0.67 – 0.87)	ABCA transporters in lipid homeostasis	rs10491178 (ABCA10), rs1860447 (ABCA9), rs3752232 (ABCA7)
0.45	0.80 (0.69 – 0.90)	Regulation of Lipid Metabolism by Peroxisome proliferator-activated receptor alpha (PPARalpha)	rs20551 (EP300), rs2305160 (NPAS2), rs9627281 (C22orf40, downstream to PPARA)
		PPARA Activates Gene Expression	
		Fatty acid, triacylglycerol, and ketone body metabolism	
0.44	0.78 (0.69 – 0.88)	Platelet Aggregation (Plug Formation)	chr4:155507965 (FGA), rs145155424 (SOS1), rs6065,CM032257 (GP1BA)
0.44	0.76 (0.64 – 0.88)	GPCR ligand binding	CM043760 (F2), rs11575580 (IL11RA), rs16940655 (CRHR1)
0.44	0.69 (0.55 – 0.82)	GPCR downstream signaling	rs11575580 (IL11RA), rs16940655 (CRHR1), rs28371560 (GHRHR)

Part V

Cytogenetic aberrations in *t(12;21)* childhood ALL

Chapter 10

Paper V - Genome-wide analysis of cytogenetic aberrations in *ETV6/RUNX1*-positive childhood acute lymphoblastic leukaemia

The following paper is a descriptive study of a sub-group of childhood ALL patients with a *t(12;21)* chromosomal translocation resulting in *ETV6/RUNX1* fusion gene. It is the most frequent chromosomal aberration in B-lineage ALL, occurring in approximately 25% of cases and it is usually associated with a more favourable prognosis. In this paper 62 patients with the *t(12;21)* translocation were analysed with Affymetrix GeneChip 500K arrays to investigate the copy number alterations (CNAs) in this group of patients.

One of the hypotheses presented by this study was that there might exist different subgroups of patients within the *t(12;21)* childhood ALL defined by their recurrent aberrations. Traditional classification approaches were not successful in identifying subgroups within the samples, possibly due to not sufficient number of patients included in the study. Therefore an integrative analysis of CNAs was applied in this study, investigating the protein-protein complexes potentially disrupted by losses of genetic material. This resulted in identification of four distinct groups of patients, characterized by disruptions in one of three identified protein-protein complexes. Details of this method are described in Chapter 7.5 and in the Supporting Material online.

The second hypothesis addressed in this study concerned the recurrent

aberrations accompanying the *ETV6/RUNX1*. The *t(12;21)* chromosomal translocation is believed to be an initiating event in development of leukaemia, however it has been hypothesized that additional genetic changes are necessary to develop leukaemia. In this study, after identifying the recurrent aberrations, we tried to investigate their sequential gain and to identify the important early leukaemogenic events, secondary to *ETV6/RUNX1*. This task was performed using a branching oncogenic tree model, resulting in a observations similar to presented in other studies of *t(12;21)* childhood ALL (Figure 2, Paper V).

Finally, the long segments of amplifications and deletions of genetic material affect the number of copies of the genes included within those segments. Therefore, we additionally examined the known microRNAs residing within those segments. microRNAs can regulate protein-coding gene targets, thus they may play an oncogenic or tumour suppressive role, depending on their transcription levels, DNA methylation and CpG content of the surrounding regions, mutations and the genes they regulate. In this study we have identified the potentially important microRNA clusters affected by CNAs and described their putative role in oncogenesis (Supporting Material online).

Supporting Material for this paper is available online from: `http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2141.2012.09083.x/suppinfo`.

bjh research paper

Genome-wide analysis of cytogenetic aberrations in *ETV6/RUNX1*-positive childhood acute lymphoblastic leukaemia

Louise Borst,[*,1] Agata Wesolowska,[*,2] Tejal Joshi [2] Rehannah Borup,[3] Finn C. Nielsen,[3] Mette K. Andersen,[4] Olafur G. Jonsson,[5] Peder S. Wehner,[6] Finn Wesenberg,[7] Britt-Marie Frost,[8] Ramneek Gupta,[2] Kjeld Schmiegelow,[1,9]

[1]Clinic for Paediatric and Adolescent Medicine, The Juliane Marie Centre, The University Hospital Rigshospitalet, Copenhagen, [2]Centre for Biological Sequence Analysis, The Technical University of Denmark, Kgs. Lyngby, [3]Centre for Genomic Medicine, The University Hospital Rigshospitalet, Copenhagen, [4]Department of Clinical Genetics, The Juliane Marie Centre, The University Hospital, Rigshospitalet, Denmark, [5]Department of Paediatrics, University Hospital, Reykjavik, Iceland, [6]Department of Paediatric Haematology and Oncology, H. C. Andersen Children's Hospital, Odense University Hospital, Odense, Denmark, [7]Department of Paediatrics, National Hospital, Oslo, Norway, [8]Institute of Paediatric, Department of Women's and Children's Health, Uppsala, Sweden, [9]The Institute of Gynaecology, Obstetrics and Paediatrics, The Faculty of Health Sciences, The University of Copenhagen, Copenhagen, Denmark

Received 20 December 2011; accepted for publication 08 February 2012

Correspondence: Kjeld Schmiegelow, Clinic for Paediatric and Youth Medicine, The Juliane Marie Centre, The University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 Copenhagen, Denmark.
E-mail: kschmiegelow@rh.dk

*Joint first authorship.

Summary

The chromosomal translocation t(12;21) resulting in the *ETV6/RUNX1* fusion gene is the most frequent structural cytogenetic abnormality among patients with childhood acute lymphoblastic leukaemia (ALL). We investigated 62 *ETV6/RUNX1*-positive childhood ALL patients by single nucleotide polymorphism array to explore acquired copy number alterations (CNAs) at diagnosis. The mean number of CNAs was 2·82 (range 0–14). Concordance with available G-band karyotyping and comparative genomic hybridization was 93%. Based on three major protein-protein complexes disrupted by these CNAs, patients could be categorized into four distinct subgroups, defined by different underlying biological mechanisms relevant to the aetiology of childhood ALL. When recurrent CNAs were evaluated by an oncogenetic tree analysis classifying their sequential order, the most common genetic aberrations (deletions of 6q, 9p, 13q and X, and gains of 10 and 21) seemed independent of each other. Finally, we identified the most common regions with recurrent gains and losses, which comprise microRNA clusters with known oncogenic or tumour-suppressive roles. The present study sheds further light on the genetic diversity of *ETV6/RUNX1*-positive childhood ALL, which may be important for understanding poor responses among this otherwise highly curable subset of ALL and lead to novel targeted treatment strategies.

Keywords: childhood acute lymphoblastic leukaemia, *ETV6/RUNX1* translocation, genome-wide screening.

Acute lymphoblastic leukaemia (ALL) is the most common cancer in children (Hjalgrim *et al*, 2003). In B-lineage childhood ALL the most frequent structural cytogenetic abnormality, present in 25% of the cases, is the chromosomal translocation t(12;21)(p13;q22) resulting in the *ETV6/RUNX1* fusion gene (Shurtleff *et al*, 1995; Forestier *et al*, 2008). The event-free survival (EFS) of childhood ALL patients after first line therapy is approximately 80% (Schmiegelow *et al*, 2010), however presence of the *ETV6/RUNX1* fusion gene is generally associated with a more favourable prognosis, with an EFS of approximately 90% (Forestier *et al*, 2008). The *ETV6/RUNX1* fusion gene is believed to

  doi:10.1111/j.1365-2141.2012.09083.x

represent an initiating event in the development of this childhood ALL subset. This initiating event, seen as early as *in utero* (Wiemels *et al*, 1999; Hjalgrim *et al*, 2002), most likely requires additional genetic changes in order to lead to leukaemia e.g. *ETV6/RUNX1* transcripts have been demonstrated in normal cord blood samples at low levels (Mori *et al*, 2002; Lausten-Thomsen *et al*, 2011) and neither *ETV6/RUNX1*-positive knock-in mouse models nor carriers of *ETV6/RUNX1*-positive pre-leukaemic cells will all develop leukaemia (Fischer *et al*, 2005; Hong *et al*, 2008). Thus, the development of *ETV6/RUNX1*-positive leukaemia requires one or more secondary genetic changes and previous studies have indeed shown additional genetic changes in the majority of *ETV6/RUNX1*-positive childhood ALL patients (Forestier *et al*, 2007; Mullighan *et al*, 2007; Kawamata *et al*, 2008; Lilljebjorn *et al*, 2007, 2010; Parker *et al*, 2008; Mullighan *et al*, 2008; van Delft *et al*, 2011).

In the present study, we applied high-resolution Affymetrix GeneChip® 500K profiling to 62 *ETV6/RUNX1*-positive childhood ALL patients to explore in detail copy number alterations (CNAs) throughout the genome. Furthermore, we investigated the inter-dependencies of the recurrent CNAs and involvement in protein-protein complexes among them, and correlated the CNAs with biologically relevant microRNAs (miRNAs) affected by the aberrations.

Methods and materials

Patients

The *ETV6/RUNX1*-positive childhood ALL patients (aged between 1 and 15 years at the time of diagnosis) from Denmark, Norway, Sweden and Iceland were diagnosed and enrolled in the Nordic Society for Paediatric Haematology and Oncology (NOPHO) treatment protocols (NOHPO ALL-92 or NOPHO ALL-2000) (Schmiegelow *et al*, 2010). A total of 133 patients were found to be positive for *ETV6/RUNX1* by fluorescent *in situ* hybridization (FISH) and/or reverse transcription polymerase chain reaction (RT-PCR) by routine investigation from 1996–2007, and diagnostic tumour DNA samples were available for 99 of these patients. Ten samples were of poor quality and/or had insufficient amount of DNA to be analysed and of the remaining samples, 62 fulfilled the quality criteria (QC) for inclusion in the bioinformatics analyses (Table 1). The study was approved by The Capital Region of Denmark Committee on Biomedical Research Ethics and The Danish Data Protection Agency.

Table 1. Patient characteristics and copy number alterations.

Characteristics	Patients
Gender	
Male	37 (59·7%)
Female	25 (40·3%)
Age, years, median (range)	4·17 (1·30–15)
WBC, $\times 10^9$/l, median (range)	11·5 (0·80–110)
Risk group	
SR	28 (45·2%)
IR	23 (37·1%)
HR	11 (17·7%)
Treatment protocol	
NOPHO ALL-92	33 (53·2%)
NOPHO ALL-2000	29 (46·8%)
Events	
Induction failure	0 (0%)
Resistant disease	0 (0%)
Relapse	4 (6·5%)
Death in remission	0 (0%)
Secondary malignancy	0 (0%)
Total events	4 (6·5%)
EFS	93·5%
CNA (mean)	
Gains > 1 Mb	0·73 (0–8)
Gains < 1 Mb	0·02 (0–1)
Losses > 1 Mb	1·24 (0–8)
Losses < 1 Mb	0·85 (0–6)
Total > 1 Mb	1·97 (0–9)
Total < 1 Mb	0·85 (0–6)
Total	2·82 (0–14)

WBC, white blood cell count; SR, standard risk; IR, intermediate risk; HR, high risk; NOPHO, Nordic Society for Paediatric Haematology and Oncology; EFS, event-free survival; CNA, copy number alteration; Mb, megabase.

G-band karyotyping and comparative genomic hybridization

G-band karyotyping was performed as part of the routine investigation on short-term cultured leukaemic cells by standard techniques. As part of a separate study, high-resolution comparative genomic hybridization (CGH) was performed for a subset of the patients (Kristensen *et al*, 2003). G-band karyotyping and/or CGH data was available in 54 (87·1%) of the included cases (Table S1).

Single nucleotide polymorphism (SNP) array analysis

DNA from bone marrow or blood samples acquired at diagnosis was extracted and purified by sodium chloride and ethanol precipitation. For SNP array analysis the DNA was processed and hybridized to Affymetrix GeneChip® Mapping 500K array set according to the manufacturer's instructions (Affymetrix, Santa Clara, CA, USA). The Affymetrix GeneChip® Mapping 500K array set comprises two arrays and is processed by two assay kits differing only in the restriction enzyme used (Affymetrix GeneChip® Mapping 250K Nsp Assay Kit and Affymetrix GeneChip® Mapping 250K Sty Assay Kit). Briefly, 250 ng DNA was digested by either Nsp I or Sty I, followed by ligation of adaptors, allowing PCR amplification of fragments in sizes from 200 to 1100 bp. PCR products were fragmented and end-labelled with biotin. Samples were subsequently hybridized to the arrays. The

arrays were washed and stained with phycoerythrin-conjugated streptavidin (SAPE) using the Affymetrix Fluidics Station® 450 and were scanned in the Affymetrix GeneChip® 2500 scanner to generate fluorescent images, as described in the Affymetrix GeneChip® protocol. Cell intensity files (CEL files) were generated in the GeneChip® operating software (GCOS) version 5.0.

Data processing

The signal intensity data based on the raw CEL files was generated using the Affymetrix Power Tools (APT) Software Package. First, for each array a QC call rate was generated using the Dynamic Model algorithm (Di *et al*, 2005). According to the manufacturer's recommendations, only samples achieving call rates $\geq$93% were used for further analysis. For the 62 samples and 200 controls fulfilling QC, genotype calls were generated using the BRLMM algorithm (http://array.mc.vanderbilt.edu/microarray/dna/brlmm.pdf). Allele-specific signals were extracted from median polish summarized and quantile normalized Perfect Match (PM) probe-intensity values. The R-GADA package (Pique-Regi *et al*, 2010) was employed to detect the recurrent CNAs. The package implements a segmentation algorithm to call CNAs based on genome alternation detection analysis (GADA). During the segmentation procedure the parameters controlling the trade-off between the sensitivity and false discovery rate (FDR) were set to achieve high specificity at the cost of sensitivity. The minimum segment length required for classification as a CNA was 8 probes, in order to exclude detection of false alterations due to extreme outliers. The 200 controls were used to exclude alterations representing the normal CNAs. Final estimation of copy number states was verified by visual examination.

Data visualization and biological correlations

The display of all the CNAs in circos format was done according to Krzywinski *et al* (2009). Recurrent CNAs were defined as gains or deletions of material occurring in the same region in at least two patients and plotted as a heatmap (Fig S1). The recurrent CNAs were then analysed further with a CRAN R package Oncotree 0·31 to determine a tree model for oncogenesis (Desper *et al*, 1999; Szabo & Boucher, 2002).

To explore functional modules that are potentially affected by the CNAs, enrichment analysis on protein-protein interaction complexes was performed using custom written Perl scripts. Only patients with losses of genetic material were included in this analysis, as these are likely to result in a direct loss of function and hence are most disruptive to the complexes. Using a curated collection of protein-protein complexes (Lage *et al*, 2007), we defined a set of complexes that categorized the patients into distinct subgroups defined by different underlying biological mechanisms. All combinations of 2–5 protein-protein complexes were tested to find a set of complexes representing the largest possible group of patients with the smallest possible overlap between different complexes. From the final collection of complex sets that best fulfilled the criteria, only those that comprised of genes highly expressed in B lymphoblasts were selected and prioritized according to leukaemia-related annotations among BioAlma terms assigned to the complexes. The miRNA annotations were based on miRBase version 16 (Griffiths-Jones, 2010) while their experimentally validated target information was collected from miRecords and miRWalk databases (Xiao *et al*, 2009; Dweep *et al*, 2011).

Results

Genomic profiling of ETV6/RUNX1-positive childhood ALL patients

The SNP array analysis revealed a total of 174 CNAs (30 were recurrent) among the 62 successfully analysed cases (whole chromosome deletions and amplifications each counted as one cytogenetic change). Of these, 129 were deletions, 77 above 1 Mb (including two whole X chromosome deletions), and 52 focal deletions below 1 Mb. A total of 45 gains were detected, of which 44 were above 1 Mb (including 18 whole chromosome gains), and only one focal amplification below 1 Mb (Fig 1 and Table S1). Since the 500K arrays do not have any probes covering the p arms of chromosomes 13, 14, 15, 21 and 22, some large amplifications of these chromosomes may actually be trisomies. The mean number of CNAs per patient was 2·82 (range 0 –14) and 13 cases (21%) did not display any CNAs (Table 1 and Table S1).

Concordance between G-band karyotyping and/or CGH and the SNP array analysis data was 93%. Thus, in 51 of the 54 cases with available data, the SNP array results either confirmed and/or added to G-band karyotyping and/or CGH (Table S1). As summarized in Fig 1, all but chromosome 17 demonstrated CNAs above 1 Mb. If focal lesions are included, all chromosomes demonstrated alterations. The most common CNAs above 1 Mb involved deletions of 12p (39%), 6q (13%), 9p (10%), 11q (10%), 13q (10%), 8p (8%) and amplifications of Xq (11%) (one female and six males), 21 (10%), 10 (5%) and Xp (5%) (two females and one male). Of the focal changes, the most common included deletions in 14q32·33 (21%), 7p14·1 (18%), 22q11·22 (13%), 14q11·2 (8%) and 7q34 (6%).

Biological correlations

Overall, neither individual aberrations nor heatmap clustering of the recurrent aberrations (Fig S1) correlated significantly with age or white blood cell count (WBC). In an attempt to classify the patients more distinctively, a systems biology approach was applied investigating the likely effects of losses of genetic material at the protein complex level. Grouping of patients in this way reflects more closely the clinical parame-

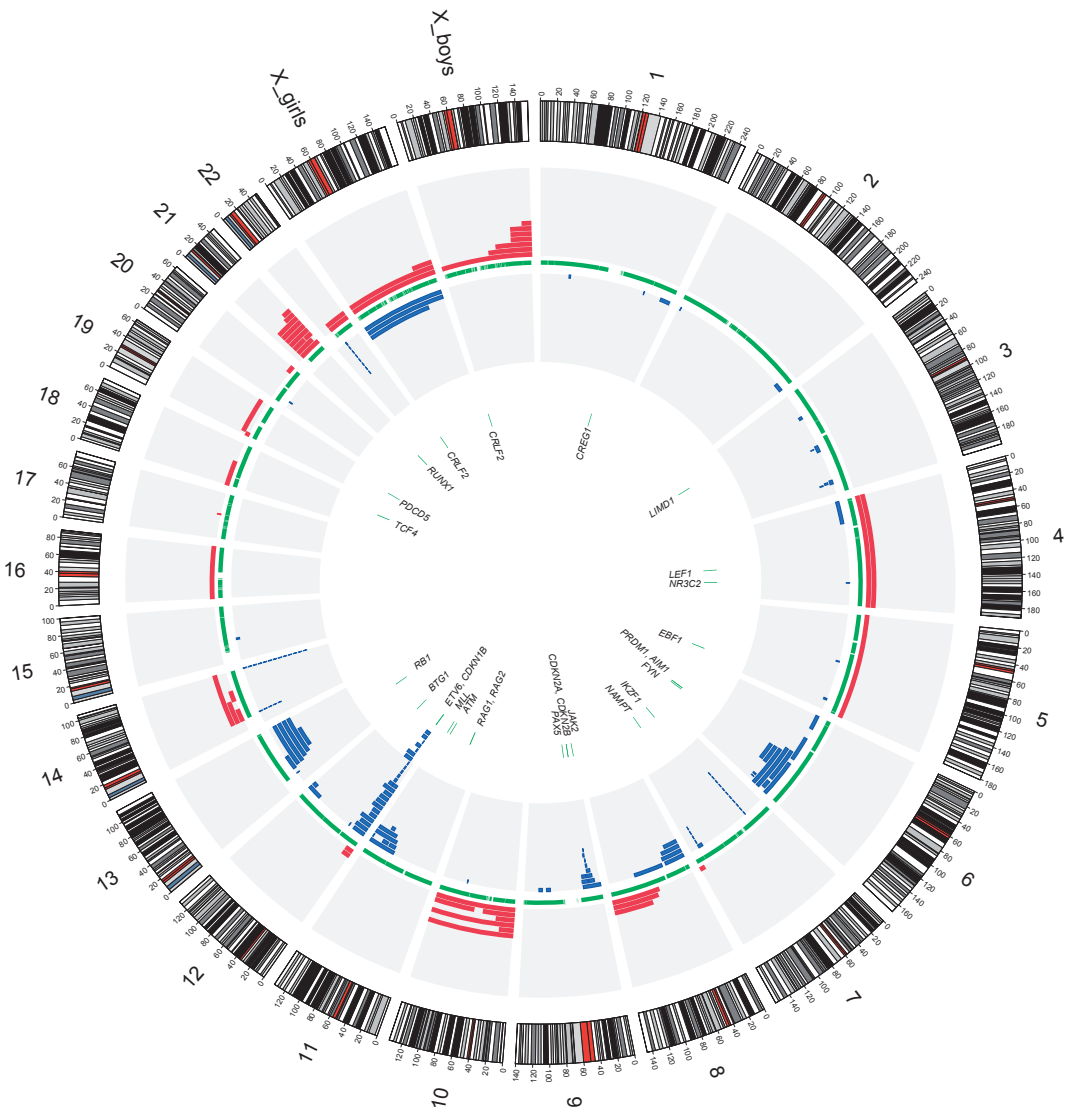


Fig 1. The recurrent cytogenetic alterations in *ETV6/RUNX1* patients plotted in Circos format (Krzywinski *et al*, 2009). The chromosomes are presented clockwise with the p arm starting from 0, followed by the centromere marked in red and by the q arm. The gains are shown in red, and the deletions in blue. The location of the probes from the 500K arrays is shown in green. The location of genes found to be important in other microarray based ALL studies is marked by green lines in the middle of the plot. Note: the alterations are stacked together; so one circle does not necessarily represent one patient.

ters and suggests different potential biological mechanisms underlying the leukaemogenesis (Data S1, Fig S2).

Next, the inter-dependency of the recurrent CNAs was investigated by an oncogenetic tree analysis to explore the sequential order of CNAs as previously done by Lilljebjorn *et al* (2010). By this approach, deletions of 12p, 14q32 and gain of 21q were classified as early events, i.e. being close to the root and acting as new roots of their own subtree (Fig 2). Importantly, even though gain of 12p (the most common additional aberration) was close to the root, it did not seem to be a required precursor for other changes. Furthermore, the most common aberrations in *ETV6/RUNX1*-positive patients (deletions of 6q, 9p, 12p, 13q, and X, and chromosome 10 and 21 amplifications) (Forestier

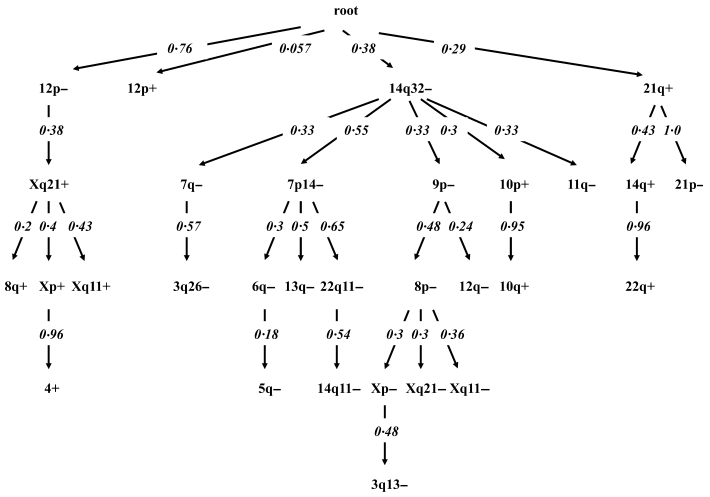


Fig 2. Oncogenetic tree. A probabilistic model of the sequential order of acquiring of the copy number alterations. The root of the tree represents the initial event - the t(12;21) chromosomal translocation. The numbers at each edge indicate the probabilities of transition along the given edge by the time of observation.

et al, 2007), seemed independent of each other, i.e. the occurrence of one did not change the likelihood of occurrence of the others. Still, 9p deletions seemed to precede 8p deletions and chromosome X deletions were rooted in 8p deletion according to the oncogenetic branching model (Fig 2). The reported estimated false-positive error rate was 0·0007 and the false-negative error rate was 0·30.

Finally, regions undergoing recurrent aberrations in the patients were also investigated for transcribed miRNAs. miRNAs can regulate protein-coding gene targets, thus they may play an oncogenic or tumour suppressive role, depending on their transcription levels, DNA methylation and CpG content of the surrounding regions, mutations and the genes they regulate (Data S2). The most frequently deleted regions, containing important miRNA clusters, were on 11q and 13q, while the most frequently amplified regions were on chromosomes 21 and X, comprising other miRNA clusters (Data S2 and Table S2).

Discussion

The present study sheds further light on the genetic diversity of *ETV6/RUNX1*-positive childhood ALL. Overall, the mean number of alterations per patient (2·82, range 0–14) is in agreement with previous studies of *ETV6/RUNX1*-positive childhood ALL patients (Forestier *et al*, 2007; Mullighan *et al*, 2007; Parker *et al*, 2008; Lilljebjorn *et al*, 2010; van Delft *et al*, 2011). Higher specificity was prioritized in attempt to reduce the rate of false positive findings, at the cost of reduced sensitivity, especially with regard to detection of focal aberrations. Furthermore, due to the size of the study, some recurrent changes may have been missed if they did not occur in at least two patients. On the other hand, the strict criteria increased the reliability of the oncogenetic tree and systems biology analyses, and is furthermore supported by the high concordance with previous results of G-band karyotyping and CGH analyses. The larger alterations: 12p, 6q, 9p, 11q, 13q and gains of chromosome 21 and Xq occurred in at least 10% of the patients, supporting previous reports (Forestier *et al*, 2007; Mullighan *et al*, 2007; Kawamata *et al*, 2008; Lilljebjorn *et al*, 2007, 2010; Parker *et al*, 2008; Mullighan *et al*, 2008; van Delft *et al*, 2011). Moreover, gain of Xq is far more common in males than in females, which supports the previous findings of Lilljebjorn *et al* (2007); (Lilljebjorn *et al*, 2010) and could imply loss of X chromosome silencing of that specific region in females leading to similar gene dosage effects in both sexes. The affected gene(s) remains to be determined.

Importantly, these non-random aberrations should be interpreted by their putative biological consequence. Here we proposed an alternative grouping of patients based on integrative analysis of recurrent CNAs and involved protein-protein interaction complexes, which may suggest different underlying biological mechanisms. This classification reflects that deletions of different genes, often from different chromosomes, may lead to disruption of the same complex resulting in a similar phenotype. In most cases the loss of a gene, evidenced through CNAs, leads to perturbation of interaction with another gene with high connectivity in the complex, and therefore is likely to have a functional effect.

To speculate on the relationship and order by which these CNAs occur, we used a branching oncogenetic tree model. The model identified deletions on 12p, 14q32 and gain of

21q as important early leukaemogenic events secondary to the *ETV6/RUNX1* translocation. Later, independent events include the most common changes in *ETV6/RUNX1*-positive childhood ALL patients, such as deletions of 6q, 9p, 13q and X and amplification of chromosome 10, suggesting these changes occur later in the development of the leukaemic clone. These implications need to be confirmed, if possible, in studies directly observing the sequential order of occurrence of CNAs. A recent study also demonstrated a similar model with data from 164 *ETV6/RUNX1*-positive childhood ALL patients (Lilljebjorn *et al*, 2010). The authors demonstrated independence between deletions on 6q, 9p and 12p, however with 6q and 9p closer to the root. Also amplification of chromosome 21 was far down a subtree rooting from 6q deletion. However, the two models cannot be compared directly because the resolution of analysed aberrations differed between the two studies. Furthermore, Anderson *et al* (2011) has also recently performed an oncogenetic tree mapping, however the study only explored pre-selected CNAs at single cell level in individual patients. Thus, the results and those from the present study are not directly comparable, but the nonlinear branching nature of the sequence of the acquired CNAs is alike.

Finally, we associated recurrent CNAs with miRNAs transcribed in those regions in order to elucidate the consequences on gene regulation of the investigated miRNAs targets. Here we observed deletions and amplifications in several chromosomal regions containing miRNAs known to be associated with tumourigenecity, leukaemia and regulation of apoptosis among others (Data S2). Thus, miRNAs may play a crucial role in childhood ALL by targeting a number of proteins and affecting various functional pathways.

Even though many clinical trials classify *ETV6/RUNX1*-positive childhood ALL as one group, their cytogenetic diversity calls for refinement in the classification of these patients, which ultimately may lead to a better understanding of their natural history (including determination of pre- and postnatal hits) (Greaves, 2006), the clinical presentation and their specific treatment requirements. Based on CNAs in protein-protein complexes, we described four groups of patients. This analysis could be further refined in the future with a larger number of samples and higher density arrays.

Acknowledgements

First of all, we are very grateful to the patients who participated in the study and their referring physicians. We also thank Hanne Maage and Charlotte Scherling for very helpful technical assistance. Acknowledgements are also made to The RH Microarray Centre/KB4111 – Department of Clinical Biochemistry – The University Hospital Rigshospitalet, Copenhagen for providing technology consultation and laboratory resources. This study has received financial support from Ministry of Health (Grant no. 2006-12103-250), The Novo Nordisk Foundation, The Danish Research Council for Health and Disease (Grant no 271-06-0278 and 271-08-0684), The Danish Childhood Cancer Foundation. Kjeld Schmiegelow holds The Danish Childhood Cancer Foundation Professorship in Paediatric Oncology.

Author contributions

LB, AW and KS designed the study, interpreted data and drafted the manuscript. LB performed the SNP array analyses in collaboration with RB and FCN, and collected patient samples and clinical data. AW performed the data processing, visualization and biological correlations. TJ performed the miRNA analyses, interpreted miRNA data and provided critical input to the manuscript. MKA was responsible for the G-band karyotyping and CGH data, while OGJ, PSW, FW and BMF provided patient samples. RG and KS contributed to the data analyses and interpretation, together with critical input to the writing of the manuscript. All authors approved the submitted and final versions of the manuscript.

Conflict of interest

The authors have no competing interests.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1. Protein-protein complexes analyses.

Data S2. MicroRNAs.

Table S1. Comparison of chromosomal alterations analysed by G-band karyotyping, CGH and SNP array analysis.

Table S2. MicroRNA clusters located in recurrent amplified and deleted regions.

Figure S1. Heatmap clustering of patients and their recurrent aberrations summarized in cytoband, chromosome arm or whole chromosome regions.

Figure S2. Protein-protein complexes.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Anderson, K., Lutz, C., van Delft, F.W., Bateman, C.M., Guo, Y., Colman, S.M., Kempski, H., Moorman, A.V., Titley, I., Swansbury, J., Kearney, L., Enver, T. & Greaves, M. (2011) Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*, **469**, 356–361.

van Delft, F.W., Horsley, S., Colman, S., Anderson, K., Bateman, C., Kempski, H., Zuna, J., Eckert, C., Saha, V., Kearney, L., Ford, A. & Greaves, M. (2011) Clonal origins of relapse in ETV6-

RUNX1 acute lymphoblastic leukemia. *Blood*, **117**, 6247–6254.

Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H. & Schaffer, A.A. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, **6**, 37–51.

Di, X., Matsuzaki, H., Webster, T.A., Hubbell, E., Liu, G., Dong, S., Bartell, D., Huang, J., Chiles, R., Yang, G., Shen, M.M., Kulp, D., Kennedy, G. C., Mei, R., Jones, K.W. & Cawley, S. (2005) Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays. *Bioinformatics*, **21**, 1958–1963.

Dweep, H., Sticht, C., Pandey, P. & Gretz, N. (2011) miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Informatics*, **44**, 839–847.

Fischer, M., Schwieger, M., Horn, S., Niebuhr, B., Ford, A., Roscher, S., Bergholz, U., Greaves, M., Lohler, J. & Stocking, C. (2005) Defining the oncogenic function of the TEL/AML1 (ETV6/RUNX1) fusion protein in a mouse model. *Oncogene*, **24**, 7579–7591.

Forestier, E., Andersen, M.K., Autio, K., Blennow, E., Borgstrom, G., Golovleva, I., Heim, S., Heinonen, K., Hovland, R., Johannsson, J.H., Kerndrup, G., Nordgren, A., Rosenquist, R., Swolin, B. & Johansson, B. (2007) Cytogenetic patterns in ETV6/RUNX1-positive pediatric B-cell precursor acute lymphoblastic leukemia: a Nordic series of 245 cases and review of the literature. *Genes, Chromosomes and Cancer*, **46**, 440–450.

Forestier, E., Heyman, M., Andersen, M.K., Autio, K., Blennow, E., Borgstrom, G., Golovleva, I., Heim, S., Heinonen, K., Hovland, R., Johannsson, J.H., Kerndrup, G., Nordgren, A., Rosenquist, R., Swolin, B. & Johansson, B. (2008) Outcome of ETV6/RUNX1-positive childhood acute lymphoblastic leukaemia in the NOPHO-ALL-1992 protocol: frequent late relapses but good overall survival. *British Journal of Haematology*, **140**, 665–672.

Greaves, M. (2006) Infection, immune responses and the aetiology of childhood leukaemia. *Nature Reviews Cancer*, **6**, 193–203.

Griffiths-Jones, S. (2010) miRBase: microRNA sequences and annotation. *Current Protocols in Bioinformatics*, **29**, 12.9.1–12.9.10.

Hjalgrim, L.L., Madsen, H.O., Melbye, M., Jorgensen, P., Christiansen, M., Andersen, M.T., Pallisgaard, N., Hokland, P., Clausen, N., Ryder, L. P., Schmiegelow, K. & Hjalgrim, H. (2002) Presence of clone-specific markers at birth in children with acute lymphoblastic leukaemia. *British Journal of Cancer*, **87**, 994–999.

Hjalgrim, L.L., Rostgaard, K., Schmiegelow, K., Soderhall, S., Kolmannskog, S., Vettenranta, K., Kristinsson, J., Clausen, N., Melbye, M., Hjalgrim, H. & Gustafsson, G. (2003) Age- and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries. *Journal of the National Cancer Institute*, **95**, 1539–1544.

Hong, D., Gupta, R., Ancliff, P., Atzberger, A., Brown, J., Soneji, S., Green, J., Colman, S., Piacibello, W., Buckle, V., Tsuzuki, S., Greaves, M. & Enver, T. (2008) Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science*, **319**, 336–339.

Kawamata, N., Ogawa, S., Zimmermann, M., Kato, M., Sanada, M., Hemminki, K., Yamatomo, G., Nannya, Y., Koehler, R., Flohr, T., Miller, C.W., Harbott, J., Ludwig, W.D., Stanulla, M., Schrappe, M., Bartram, C.R. & Koeffler, H.P. (2008) Molecular allelokaryotyping of pediatric acute lymphoblastic leukemias by high-resolution single nucleotide polymorphism oligonucleotide genomic microarray. *Blood*, **111**, 776–784.

Kristensen, T.D., Wesenberg, F., Jonsson, O.G., Carlsen, N.T., Forestier, E., Kirchhoff, M., Lundsteen, C. & Schmiegelow, K. (2003) High-resolution comparative genomic hybridisation yields a high detection rate of chromosomal aberrations in childhood acute lymphoblastic leukaemia. *European Journal of Haematology*, **70**, 363–372.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. & Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Research*, **19**, 1639–1645.

Lage, K., Karlberg, E.O., Storling, Z.M., Olason, P. I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tumer, Z., Pociot, F., Tommerup, N., Moreau, Y. & Brunak, S. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, **25**, 309–316.

Lausten-Thomsen, U., Madsen, H.O., Vestergaard, T.R., Hjalgrim, H., Nersting, J. & Schmiegelow, K. (2011) Prevalence of t(12;21)[ETV6-RUNX1]-positive cells in healthy neonates. *Blood*, **117**, 186–189.

Lilljebjorn, H., Heidenblad, M., Nilsson, B., Lassen, C., Horvat, A., Heldrup, J., Behrendtz, M., Johansson, B., Andersson, A. & Fioretos, T. (2007) Combined high-resolution array-based comparative genomic hybridization and expression profiling of ETV6/RUNX1-positive acute lymphoblastic leukemias reveal a high incidence of cryptic Xq duplications and identify several putative target genes within the commonly gained region. *Leukemia*, **21**, 2137–2144.

Lilljebjorn, H., Soneson, C., Andersson, A., Heldrup, J., Behrendtz, M., Kawamata, N., Ogawa, S., Koeffler, H.P., Mitelman, F., Johansson, B., Fontes, M. & Fioretos, T. (2010) The correlation pattern of acquired copy number changes in 164 ETV6/RUNX1-positive childhood acute lymphoblastic leukemias. *Human Molecular Genetics*, **19**, 3150–3158.

Mori, H., Colman, S.M., Xiao, Z., Ford, A.M., Healy, L.E., Donaldson, C., Hows, J.M., Navarrete, C. & Greaves, M. (2002) Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proceedings of the National Academy of Sciences USA*, **99**, 8242–8247.

Mullighan, C.G., Goorha, S., Radtke, I., Miller, C. B., Coustan-Smith, E., Dalton, J.D., Girtman, K., Mathew, S., Ma, J., Pounds, S.B., Su, X., Pui, C.H., Relling, M.V., Evans, W.E., Shurtleff, S.A. & Downing, J.R. (2007) Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature*, **446**, 758–764.

Mullighan, C.G., Miller, C.B., Radtke, I., Phillips, L.A., Dalton, J., Ma, J., White, D., Hughes, T.P., Le Beau, M.M., Pui, C.H., Relling, M.V., Shurtleff, S.A. & Downing, J.R. (2008) BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature*, **453**, 110–114.

Parker, H., An, Q., Barber, K., Case, M., Davies, T., Konn, Z., Stewart, A., Wright, S., Griffiths, M., Ross, F.M., Moorman, A.V., Hall, A.G., Irving, J.A., Harrison, C.J. & Strefford, J.C. (2008) The complex genomic profile of ETV6-RUNX1 positive acute lymphoblastic leukemia highlights a recurrent deletion of TBL1XR1. *Genes, Chromosomes and Cancer*, **47**, 1118–1125.

Pique-Regi, R., Caceres, A. & Gonzalez, J.R. (2010) R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics*, **11**, 380.

Schmiegelow, K., Forestier, E., Hellebostad, M., Heyman, M., Kristinsson, J., Soderhall, S. & Taskinen, M. (2010) Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia*, **24**, 345–354.

Shurtleff, S.A., Buijs, A., Behm, F.G., Rubnitz, J.E., Raimondi, S.C., Hancock, M.L., Chan, G.C., Pui, C.H., Grosveld, G. & Downing, J.R. (1995) TEL/AML1 fusion resulting from a cryptic t (12;21) is the most common genetic lesion in pediatric ALL and defines a subgroup of patients with an excellent prognosis. *Leukemia*, **9**, 1985–1989.

Szabo, A. & Boucher, K. (2002) Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical Biosciences*, **176**, 219–236.

Wiemels, J.L., Cazzaniga, G., Daniotti, M., Eden, O.B., Addison, G.M., Masera, G., Saha, V., Biondi, A. & Greaves, M.F. (1999) Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet*, **354**, 1499–1503.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. & Li, T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, **37**, D105–D110.

Part VI

Epilogue

Chapter 11

Summary and perspectives

This thesis presents the current state-of-art in the field of functional genomic variations and introduces childhood acute lymphoblastic leukaemia as an excellent model for studying pharmacogenomics. The main thrust of this thesis concentrates on investigating genomic variations in context of inter-individual differences in patients' treatment response (Paper III), susceptibility to adverse side effects (Paper IV) and identifying subgroups of patients with potentially different disease aetiology (Paper V). These investigations are accompanied by discussion of available genotyping technologies (Chapter 5 and Chapter 6.2) and introduction of a novel, cost-effective method of genotyping by multiplexed targeted sequencing (Paper II). Chapter 4 provides a brief overview of predicting effects of SNPs on their transcript, subsequent prediction of deleteriousness of a given variant, accompanied by comparison of mutability and deleteriousness of amino acid changes (Paper I) and finally identifying more specific molecular mechanisms affected by the variant by applying an ensemble of prediction tools for proteins. The presented strategies to discover associations to phenotypes of interest include associations of single SNPs as well as sets of SNPs grouped by functional modules in which they are likely to operate (Chapter 7). One of the methods applied in Paper III (described in Chapter 7.2) facilitates including the rare variant genotypes in the analysis to investigate accumulation of variants in certain genomic regions or protein-protein complexes, and a pathway analysis (described in Chapter 7.3) applied in Papers III and IV allows more robust analysis identifying non-linear interactions between genotypes and pinpointing well-defined biological mechanisms contributing to the phenotype. Furthermore, a discussion on personal genomes (Chapter 1.4) is followed by a suggestion of a simple way of assessing individual disease risk based on overrepresentation of genomic variants in disease-associated gene sets (Chapter 7.4). Finally, integrative analysis of acquired copy number

alterations presented in Chapter 7.5 and Paper V introduces a way of elucidating different biological mechanisms underlying the leukaemogenesis by identifying disruptions of functional protein-protein complexes.

Overall, the work presented in this thesis provides a proof-of-concept that large-scale hypothesis-driven studies are an emerging alternative to data-driven GWA studies. Such a study design facilitates interpretation of results and integration of data for analyses of effects mediated via various genes acting in the same pathway or protein-protein complex. Several other studies on clinical aspects of treatment in childhood ALL have been planned and approved by the NOPHO society, including studies on life-threatening toxicities and metabolism of specific drugs.

In the remaining part of this thesis I will conclude with some final remarks on the perspectives of the field of pharmacogenomics and personalized medicine.

11.1 Functional variations

The work presented in this thesis concerns mostly germline genomic variation (except for Paper V), however when studying cancer one should not forget that the observed clinical phenotypes are a result of a complex interplay between the host and tumour genetic factors. Since leukaemia is generally a chemosensitive cancer [24], the hypothesis presented in Chapter 7 stated that majority of leukaemia relapses are caused by insufficient systemic exposure to certain drugs, rather than cancer chemoresistance. The results of the study (Paper III) are quite promising, however they still fail to explain all the incidences of relapse. In order to fully understand this process ideally one would examine matched host and tumour samples. It is also possible that other factors beyond point mutations and polymorphisms may contribute to the phenotype, including copy number variation, methylation patterns, as well as expression patterns of genes, proteins and non-coding RNAs. Clearly, relapse can be caused by different molecular mechanisms and is therefore impossible to be explained by a single variant in the genome. On the other hand, MRD levels after remission induction therapy are likely to be sensitive to changes in treatment protocols and measurement methods. Therefore integrative variant analysis provides a way of capturing those differences and identifying common affected mechanisms of action, defined for instance by biological pathways. The study on infection susceptibility, even though conducted on much smaller group of patients, yields seemingly more accurate classifications, possibly due to clear definition of the phenotype and considerate contribution of genetics. This reflects the fact that one of the keys to a successful association study is a clear definition of a phenotype and leads us to believe that future studies planned by the NOPHO society concerning clearance of specific drugs or susceptibilities to particular toxicities may yield

even more interesting results.
One of the major difficulties in studies investigating impact of genomic variation on observed phenotypes is the gap in our understanding of non-coding and synonymous variations. Our lack of understanding of the potential functionality of those variants does not imply unimportance and eventually those types of variations have to be addressed in the analyses. Furthermore, there is increasing evidence that rare variations can have a considerate contribution to a variety of phenotypes and so far the only efforts to include this type of variants in analysis is to investigate whether particular genomic regions are enriched in rare variations, without trying to interpret and understand their actual functionality. Finally, one of the limiting factors in genomic variation analyses might be the incompleteness of the reference human genome, which for most parts of the genome likely reflects the DNA of any individual, but might differ significantly in highly variable or repetitive regions. A recent study by MacArtur *et al.* [77] reports that any individual carries approximately 100 loss-of-function variants, resulting in approximately 20 genes being completely inactivated which leads to great inter-individual variability of gene content. Clearly, it is easy to imagine an opposite situation in which all individuals probably have a unique copy of a number of genes. This is an obvious obstacle in the current methods for NGS data analysis where the sequencing reads are mapped back to the reference genome, since certain regions of the genome are not yet characterized. When doing SNP calling analyses, the normal procedure it to compare the observed sequence with the reference genome, which introduces the assumption of the contents of the reference sequence being ”normal” while any differences being ”alternative alleles”, even if the ”alternative” allele is in fact the prevalent one in general population. In future, application of single molecule sequencing could help overcome some of those problems, however this technology is still burdened with very high error rates and until this methodology improves one can attempt to create alternative reference genomes by population specific genome assemblies.

11.2 Personalized medicine

The field of personalized medicine is rather far from the vision of applying drugs specifically tailored to patients genetic profiles, however it is getting much closer to optimizing the drug dosing based on an individual's DNA. With the exponential decline in sequencing costs and constant improvements in the technology, NGS is making its way into the clinic to assist in the clinical interpretation of patient's genetic profiles. The findings of pharmacogenetics studies are starting to be introduced into clinical settings, and up to date 118 drugs contain labels approved by the Food and Drug Administration (FDA) agency recommending genotype-specific dosing or at least recognizing the influence of genetic variation on drug response or safety (source: `www.fda.gov`). Furthermore, genomic data has been shown

to improve the disease classifications and to help redefining the disease phenotypes as well as the therapeutic strategies. For instance, whole-genome expression profiles have been utilized to identify a new subclass of Burkitt 's lymphoma from diffuse B-cell lymphoma without prior knowledge of the classes [32], while estrogen receptor status can be used to determine the best treatment options in breast cancer [126]. Integration of the genomic information together with its derivatives, such as the transcriptome, proteome and metabolome seems crucial to investigate the whole landscape of a disease. Recent story of Michael Snyder's personal omics [23] shows that indeed this approach can be successful in predicting the disease risks and monitoring of the biological changes between healthy and diseased states. His integrative personal omics profile indicated a high risk of developing type 2 diabetes, despite lack of family history of the disease. This finding motivated him to monitor the glucose levels and indeed during the time of the study he discovered that he had developed the disease, however due to detection in early stage few changes in life-style were sufficient to prevent the further development of the disease. Since the content of our DNA is stable during lifetime, we are in principle able to predict the disease risk and susceptibility at any time in our life and therefore we may hope to observe a shift in medical care expenses from disease treatment to disease prevention in the future. Before that happens, several ethical issues have to be resolved including the issue whether to inform patients about the risks of untreatable diseases or whether patients should be able to decide about this themselves. As an example, when James Watson had his genome sequenced, the only region he refused to reveal was the *APOE* gene containing a known variant significantly increasing the susceptibility to Alzheimer's disease. Finally, the issue of privacy of the genomic data has to be addressed, since sharing of the data could have potential economical or social aspects including potential increase in health insurance cost or difficulties in finding a job due to predicted high disease risk.

Regardless of the technological and ethical issues, genomic data is the driving force behind personalized medicine. Personalized medicine is not likely to change the traditional medicine but instead it can optimize it to individual patient's needs increasing the efficacy and at the same time reducing the side effects of the treatment.

Bibliography

[1] Adzhubei I., Schmidt S., Peshkin L., Ramensky V., Gerasimova A., et al. (2010). A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249. 24

[2] Aitman T., Dong R., Vyse T., Norsworthy P., Johnson M., et al. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078):851–855. 64

[3] Allen H., Estrada K., Lettre G., Berndt S., Weedon M., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838. 7

[4] Almal S. and Padh H. (2011). Implications of gene copy-number variation in health and diseases. *Journal of human genetics*, 57(1):6–13. 3

[5] Altshuler D., Lander E., Ambrogio L., Bloom T., Cibulskis K., et al. (2010). A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061–1073. 3

[6] Ashburner M., Ball C., Blake J., Botstein D., Butler H., et al. (2000). Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25. 62

[7] Ashley E., Butte A., Wheeler M., Chen R., Klein T., et al. (2010). Clinical assessment incorporating a personal genome. *The Lancet*, 375(9725):1525–1535. 62, 64

[8] Bader G., Donaldson I., Wolting C., Ouellette B., Pawson T., et al. (2001). BIND—the biomolecular interaction network database. *Nucleic acids research*, 29(1):242–245. 43

[9] Bairoch A., Apweiler R., Wu C., Barker W., Boeckmann B., et al. (2005). The universal protein resource (UniProt). *Nucleic acids research*, 33(suppl 1):D154–D159. 43

[10] Bansal V., Libiger O., Torkamani A., and Schork N. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics*, 11(11):773–785. 58, 59

[11] Begg S., Vos T., Barker B., Stevenson C., Stanley L., et al. (2007). *Burden of disease and injury in Australia, 2003*. Australian Institute of Health and Welfare AIHW. 62, 63

[12] Bentley D., Balasubramanian S., Swerdlow H., Smith G., Milton J., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59. 39

[13] Bodmer W. and Bonilla C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, 40(6):695–701. 58

[14] Borst L., Buchard A., Rosthøj S., Wesolowska A., Wehner P., et al. (2012). Gene dose effects of GSTM1, GSTT1 and GSTP1 polymorphisms on outcome in childhood acute lymphoblastic leukemia. *Journal of Pediatric Hematology/Oncology*, 34(1):38. 40

[15] Bosch T., Meijerman I., Beijnen J., and Schellens J. (2006). Genetic polymorphisms of drug-metabolising enzymes and drug transporters in the chemotherapeutic treatment of cancer. *Clinical pharmacokinetics*, 45(3):253–285. 16

[16] Brockman W., Alvarez P., Young S., Garber M., Giannoukos G., et al. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Research*, 18(5):763–770. 37

[17] Broome J. (1981). L-Asparaginase: discovery and development as a tumor-inhibitory agent. *Cancer treatment reports*, 65:111. 18

[18] Brown K. and Jurisica I. (2005). Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082. 43

[19] Burton P., Clayton D., Cardon L., Craddock N., Deloukas P., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678. 6

[20] Carnevali P., Baccash J., Halpern A., Nazarenko I., Nilsen G., et al. (2012). Computational techniques for human genome resequencing using mated gapped reads. *Journal of Computational Biology*, 19(3):279–292. 38

[21] Chabner B. (1996). Cytidine analogues. *Cancer chemotherapy and biotherapy: principles and practice. 2nd ed. Philadelphia: Lippincott-Raven*, pages 213–33. 19

[22] Chatr-Aryamontri A., Ceol A., Palazzi L., Nardelli G., Schneider M., et al. (2007). MINT: the Molecular INTeraction database. *Nucleic acids research*, 35(suppl 1):D572–D574. 43

[23] Chen R., Mias G., Li-Pook-Than J., Jiang L., Lam H., et al. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148(6):1293–1307. 126

[24] Cheok M. and Evans W. (2006). Acute lymphoblastic leukaemia: a model for the pharmacogenomics of cancer therapy. *Nature Reviews Cancer*, 6(2):117–129. 16, 124

[25] Church G. (2005). The personal genome project. *Molecular Systems Biology*, 1(1). 10

[26] Clarke G., Anderson C., Pettersson F., Cardon L., Morris A., et al. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133. 5

[27] Cook Jr E. and Scherer S. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(7215):919–923. 64

[28] Cooper D., Stenson P., and Chuzhanova N. (2006). The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Current protocols in bioinformatics*, pages 1–13. 24, 43

[29] Cooper G. and Shendure J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9):628–640. 6, 24

[30] Craddock N., Hurles M., Cardin N., Pearson R., Plagnol V., et al. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720. 64

[31] Croft D., O'Kelly G., Wu G., Haw R., Gillespie M., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl 1):D691–D697. 43, 60, 62

[32] Dave S., Fu K., Wright G., Lam L., Kluin P., et al. (2006). Molecular diagnosis of Burkitt's lymphoma. *New England Journal of Medicine*, 354(23):2431–2442. 126

[33] Davidsen M., Dalhoff K., and Schmiegelow K. (2008). Pharmacogenetics influence treatment efficacy in childhood acute lymphoblastic leukemia. *Journal of pediatric hematology/oncology*, 30(11):831–849. 14, 16, 17, 41

[34] Davies S., Borowitz M., Rosner G., Ritz K., Devidas M., et al. (2008). Pharmacogenetics of minimal residual disease response in children with B-precursor acute lymphoblastic leukemia: a report from the Children's Oncology Group. *Blood*, 111(6):2984–2990. 71

[35] Davis A., Murphy C., Saraceni-Richards C., Rosenstein M., Wiegers T., et al. (2009). Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical–gene–disease networks. *Nucleic acids research*, 37(suppl 1):D786–D792. 42, 43

[36] Dennis Jr G., Sherman B., Hosack D., Yang J., Gao W., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3. 62, 64

[37] DePristo M., Banks E., Poplin R., Garimella K., Maguire J., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498. 36

[38] Du P., Feng G., Flatow J., Song J., Holko M., et al. (2009). From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, 25(12):i63–i68. 62, 64

[39] Eden T. (2010). Aetiology of childhood leukaemia. *Cancer treatment reviews*, 36(4):286–297. 11

[40] Estlin E., Yule S., and Lowis S. (2001). Consolidation therapy for childhood acute lymphoblastic leukaemia: Clinical and cellular pharmacology of cytosine arabinoside, epipodophyllotoxins and cyclophosphamide. *Cancer treatment reviews*, 27(6):339. 20

[41] Ewing B. and Green P. (1998). Base-calling of automated sequencer traces using-Phred. II. error probabilities. *Genome research*, 8(3):186–194. 35

[42] Forbes S., Bindal N., Bamford S., Cole C., Kok C., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research*, 39(suppl 1):D945–D950. 43

[43] Fujimoto A., Nakagawa H., Hosono N., Nakano K., Abe T., et al. (2010). Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. *Nature genetics*, 42(11):931–936. 64

[44] Garrison E. and Marth G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*. 38

[45] Gibbs R., Belmont J., Hardenbol P., Willis T., Yu F., et al. (2003). The international HapMap project. *Nature*, 426(6968):789–796. 4

[46] Gnirke A., Melnikov A., Maguire J., Rogov P., LeProust E., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature biotechnology*, 27(2):182–189. 44

[47] Gottesman M., Fojo T., Bates S., et al. (2002). Multidrug resistance in cancer: role of ATP-dependent transporters. *Nature Reviews Cancer*, 2(1):48–58. 16

[48] Gregers J., Christensen I., Dalhoff K., Lausen B., Schroeder H., et al. (2010). The association of reduced folate carrier 80G> A polymorphism to outcome in childhood acute lymphoblastic leukemia interacts with chromosome 21 copy number. *Blood*, 115(23):4671–4677. 18

[49] Güldener U., Münsterkötter M., Oesterheld M., Pagel P., Ruepp A., et al. (2006). MPact: the MIPS protein interaction resource on yeast. *Nucleic acids research*, 34(suppl 1):D436–D441. 43

[50] Hamosh A., Scott A., Amberger J., Bocchini C., and McKusick V. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517. 5, 43, 62

[51] Hande K. et al. (1998). Etoposide: four decades of development of a topoisomerase II inhibitor. *European journal of cancer (Oxford, England: 1990)*, 34(10):1514. 20

[52] Hedeland R., Hvidt K., Nersting J., Rosthøj S., Dalhoff K., et al. (2010). DNA incorporation of 6-thioguanine nucleotides during maintenance therapy of childhood acute lymphoblastic leukaemia and non-Hodgkin lymphoma. *Cancer chemotherapy and pharmacology*, 66(3):485–491. 19

[53] Hermjakob H., Montecchi-Palazzi L., Lewington C., Mudali S., Kerrien S., et al. (2004). IntAct: an open source molecular interaction database. *Nucleic acids research*, 32(suppl 1):D452–D455. 43

[54] Hewett M., Oliver D., Rubin D., Easton K., Stuart J., et al. (2002). PharmGKB: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165. 15, 41, 43, 60

[55] Hiard S., Charlier C., Coppieters W., Georges M., and Baurain D. (2010). Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic acids research*, 38(suppl 1):D640–D651. 42, 43

[56] Hindorff LA M.J.J.H.H.P.K.A.M.T. MacArthur J (????). A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. 5, 43

[57] Hjalgrim L., Rostgaard K., Schmiegelow K., Söderhäll S., Kolmannskog S., et al. (2003). Age-and sex-specific incidence of childhood leukemia by immunophenotype in the Nordic countries. *Journal of the National Cancer Institute*, 95(20):1539–1544. 11

[58] Hubbard T., Barker D., Birney E., Cameron G., Chen Y., et al. (2002). The Ensembl genome database project. *Nucleic acids research*, 30(1):38–41. 24, 42

[59] Imai K., Kricka L., and Fortina P. (2011). Concordance study of 3 direct-to-consumer genetic-testing services. *Clinical chemistry*, 57(3):518–521. 10

[60] Iqbal Z., Caccamo M., Turner I., Flicek P., and McVean G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*. 38

[61] Johannsen W. (1911). The genotype conception of heredity. *American Naturalist*, pages 129–159. 4

[62] Johansen C., Wang J., Lanktree M., Cao H., McIntyre A., et al. (2010). Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics*, 42(8):684–687. 64

[63] Jordan M. and Wilson L. (2004). Microtubules as a target for anticancer drugs. *Nature Reviews Cancer*, 4(4):253–265. 18

[64] Kahvejian A., Quackenbush J., and Thompson J. (2008). What would you do if you could sequence everything? *Nature biotechnology*, 26(10):1125–1133. 8

[65] Kimchi-Sarfaty C., Oh J., Kim I., Sauna Z., Calcagno A., et al. (2007). A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811):525–528. 25

[66] Komar A. (2007). SNPs, silent but not invisible. *DNA*, 5:3. 25

[67] Lage K., Hansen N., Karlberg E., Eklund A., Roque F., et al. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences*, 105(52):20870–20875. 62

[68] Lage K., Karlberg E., Størling Z., Olason P., Pedersen A., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature biotechnology*, 25(3):309–316. 42, 43, 62

[69] Lamba J. (2009). Genetic factors influencing cytarabine therapy. *Pharmacogenomics*, 10(10):1657–1674. 19

[70] Lander E., Linton L., Birren B., Nusbaum C., Zody M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 3, 4

[71] Langmead B., Trapnell C., Pop M., Salzberg S., et al. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25. 36

[72] Lee W., Yue P., and Zhang Z. (2009). Analytical methods for inferring functional effects of single base pair substitutions in human cancers. *Human genetics*, 126(4):481–498. 24

[73] Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760. 36

[74] Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. 38

[75] Lund B., Åsberg A., Heyman M., Kanerva J., Harila-Saari A., et al. (2011). Risk factors for treatment related mortality in childhood acute lymphoblastic leukaemia. *Pediatric blood & cancer*, 56(4):551–559. 93

[76] Lunter G. and Goodson M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research*, 21(6):936–939. 36

[77] MacArthur D., Balasubramanian S., Frankish A., Huang N., Morris J., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828. 125

[78] Maher B. (2008). The case of the missing heritability. *Nature*, 456(7218):18–21. 58

[79] Malhotra A. (2010). The State of Pharmacogenetics Customizing Treatments. *Psychiatric Times*, 27(4):1. 15

[80] McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303. 37, 38

[81] McLaren W., Pritchard B., Rios D., Chen Y., Flicek P., et al. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070. 23, 42

[82] Mendel G. (1865). Experiments in plant hybridization (1865). *Read at the February*, 8:3–47. 4

[83] Metzker M. (2009). Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31–46. 7

[84] Mikkelsen T., Thorn C., Yang J., Ulrich C., French D., et al. (2011). PharmGKB summary: methotrexate pathway. *Pharmacogenetics and genomics*, 21(10):679–686. 18

[85] Minotti G., Menna P., Salvatorelli E., Cairo G., and Gianni L. (2004). Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacological reviews*, 56(2):185–229. 18

[86] Nelson M., Wegmann D., Ehm M., Kessner D., Jean P., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104. 58

[87] Ng P. and Henikoff S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814. 24

[88] Ng S., Buckingham K., Lee C., Bigham A., Tabor H., et al. (2009). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35. 9, 23

[89] Ng S., Turner E., Robertson P., Flygare S., Bigham A., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276. 9, 23

[90] Nielsen R., Paul J., Albrechtsen A., and Song Y. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451. 37

[91] Nijman I., Mokry M., van Boxtel R., Toonen P., de Bruijn E., et al. (2010). Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nature methods*, 7(11):913–915. 44

[92] Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., et al. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 27(1):29–34. 43, 62

[93] Olshen A., Venkatraman E., Lucito R., and Wigler M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572. 38

[94] Ong V., Liem N., Schmid M., Verrills N., Papa R., et al. (2008). A role for altered microtubule polymer levels in vincristine resistance of childhood acute lymphoblastic leukemia xenografts. *Journal of Pharmacology and Experimental Therapeutics*, 324(2):434–442. 18

[95] Osborne J., Flatow J., Holko M., Lin S., Kibbe W., et al. (2009). Annotating the human genome with Disease Ontology. *BMC genomics*, 10(Suppl 1):S6. 62, 64

[96] Pagel P., Kovac S., Oesterheld M., Brauner B., Dunger-Kaltenbach I., et al. (2005). The MIPS mammalian protein–protein interaction database. *Bioinformatics*, 21(6):832–834. 43

[97] Peri S., Navarro J., Amanchy R., Kristiansen T., Jonnalagadda C., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10):2363–2371. 43

[98] Pique-Regi R., Monso-Varona J., Ortega A., Seeger R., Triche T., et al. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 24(3):309–318. 38

[99] Pui C., Carroll W., Meshinchi S., and Arceci R. (2011). Biology, risk stratification, and therapy of pediatric acute leukemias: an update. *Journal of Clinical Oncology*, 29(5):551–565. 13

[100] Pui C., Relling M., and Evans W. (2002). Role of pharmacogenomics and pharmacodynamics in the treatment of acute lymphoblastic leukaemia. *Best Practice & Research Clinical Haematology*, 15(4):741–756. 18, 19

[101] Ramsey L., Bruun G., Yang W., Treviño L., Vattathil S., et al. (2012). Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. *Genome Research*, 22(1):1–8. 58

[102] Rasmussen M., Guo X., Wang Y., Lohmueller K., Rasmussen S., et al. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, 334(6052):94–98. 10, 62

[103] Rasmussen M., Li Y., Lindgreen S., Pedersen J., Albrechtsen A., et al. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*, 463(7282):757–762. 10

[104] Rebhan M., Chalifa-Caspi V., Prilusky J., and Lancet D. (1998). GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656–664. 62

[105] Redon R., Ishikawa S., Fitch K., Feuk L., Perry G., et al. (2006). Global variation in copy number in the human genome. *nature*, 444(7118):444–454. 3

[106] Reik W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432. 4

[107] Relling M., Gardner E., Sandborn W., Schmiegelow K., Pui C., et al. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clinical Pharmacology & Therapeutics*, 89(3):387–391. 19

[108] Relling M., Hancock M., Rivera G., Sandlund J., Ribeiro R., et al. (1999). Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus. *Journal of the National Cancer Institute*, 91(23):2001–2008. 19, 20

[109] Rocha J., Cheng C., Liu W., Kishi S., Das S., et al. (2005). Pharmacogenetics of outcome in children with acute lymphoblastic leukemia. *Blood*, 105(12):4752–4758. 71

[110] Rosenbloom K., Dreszer T., Long J., Malladi V., Sloan C., et al. (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic acids research*, 40(D1):D912–D917. 25

[111] Ruepp A., Brauner B., Dunger-Kaltenbach I., Frishman G., Montrone C., et al. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(suppl 1):D646–D650. 43

[112] Sadee W. and Dai Z. (2005). Pharmacogenetics/genomics and personalized medicine. *Human molecular genetics*, 14(suppl 2):R207–R214. 15

[113] Salwinski L., Miller C., Smith A., Pettit F., Bowie J., et al. (2004). The database of interacting proteins: 2004 update. *Nucleic acids research*, 32(suppl 1):D449–D451. 43

[114] Schlessinger A., Matsson P., Shima J., Pieper U., Yee S., et al. (2010). Comparison of human solute carriers. *Protein Science*, 19(3):412–428. 16

[115] Schmiegelow G.G. K. (2005). Acute Lymphoblastic Leukemia. In S.M.C.H. Voute PA Barrett A, editor, *Cancer in children: clinical management*, chapter 12, pages 138–69. Oxford University Press, 5th edition. 13

[116] Schmiegelow K., Forestier E., Hellebostad M., Heyman M., Kristinsson J., et al. (2009). Long-term results of NOPHO ALL-92 and ALL-2000 studies of childhood acute lymphoblastic leukemia. *Leukemia*, 24(2):345–354. 13, 19

[117] Schmiegelow K., Forestier E., Kristinsson J., Söderhäll S., Vettenranta K., et al. (2008). Thiopurine methyltransferase activity is related to the risk of relapse of childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study. *Leukemia*, 23(3):557–564. 14

[118] Sherry S., Ward M., Kholodov M., Baker J., Phan L., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311. 4, 42, 43

[119] Siva N. (2008). 1000 Genomes project. *Nature biotechnology*, 26(3):256–256. 4

[120] Stanulla M., Schaeffeler E., Flohr T., Cario G., Schrauder A., et al. (2005). Thiopurine methyltransferase (TPMT) genotype and early treatment response to mercaptopurine in childhood acute lymphoblastic leukemia. *JAMA: the journal of the American Medical Association*, 293(12):1485–1489. 71

[121] Stark C., Breitkreutz B., Reguly T., Boucher L., Breitkreutz A., et al. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539. 43

[122] Subramanian A., Tamayo P., Mootha V., Mukherjee S., Ebert B., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550. 60, 62, 64

[123] Teitell M. and Pandolfi P. (2009). Molecular genetics of acute lymphoblastic leukemia. *Annual Review of Pathological Mechanical Disease*, 4:175–198. 12

[124] Thorn C., Oshiro C., Marsh S., Hernandez-Boussard T., McLeod H., et al. (2011). Doxorubicin pathways: pharmacodynamics and adverse effects. *Pharmacogenetics and genomics*, 21(7):440. 18

[125] Tissing W., Meijerink J., Den Boer M., and Pieters R. (2003). Molecular determinants of glucocorticoid sensitivity and resistance in acute lymphoblastic leukemia. *Leukemia*, 17(1):17–25. 17

[126] van't Veer L., Dai H., Van De Vijver M., He Y., Hart A., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536. 126

[127] Venter J., Adams M., Myers E., Li P., Mural R., et al. (2001). The sequence of the human genome. *Science Signalling*, 291(5507):1304. 3, 4

[128] Vernot B., Stergachis A., Maurano M., Vierstra J., Neph S., et al. (2012). Personal and population genomics of human regulatory variation. *Genome Research*, 22(9):1689–1697. 25

[129] von Bubnoff A. (2008). Next-generation sequencing: the race is on. *Cell*, 132(5):721–723. 8

[130] Wang K., Li M., and Hakonarson H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854. 60

[131] Wishart D., Knox C., Guo A., Cheng D., Shrivastava S., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906. 15, 19, 41, 43

[132] Yang J., Bogni A., Schuetz E., Ratain M., Eileen Dolan M., et al. (2009). Etoposide pathway. *Pharmacogenetics and Genomics*, 19(7):552. 20

[133] Yang J., Cheng C., Devidas M., Cao X., Campana D., et al. (2012). Genome-wide association study identifies germline polymorphisms associated with relapse of childhood acute lymphoblastic leukemia. *Blood*. 71

[134] Yang J., Cheng C., Yang W., Pei D., Cao X., et al. (2009). Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *JAMA: the journal of the American Medical Association*, 301(4):393–403. 71

[135] Yoshiura K., Kinoshita A., Ishida T., Ninokata A., Ishikawa T., et al. (2006). A SNP in the ABCC11 gene is the determinant of human earwax type. *Nature genetics*, 38(3):324–330. 5

[136] Zawistowski M., Gopalakrishnan S., Ding J., Li Y., Grimm S., et al. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *American journal of human genetics*, 87(5):604. 58

[137] Zaza G., Cheok M., Krynetskaia N., Thorn C., Stocco G., et al. (2010). Thiopurine pathway. *Pharmacogenetics and genomics*, 20(9):573. 19

[138] Zondervan K. and Cardon L. (2007). Designing candidate gene and genome-wide case–control association studies. *Nature protocols*, 2(10):2492–2501. 7